(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号
特開2003－131917
（P2003－131917A）
(43)公開日　平成15年5月9日(2003.5.9)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
G06F	12/00	533	G06F	12/00	533A　5B065
		545			545A　5B082
	3/06	304		3/06	304F
					304P

審査請求　未請求　請求項の数8　OL　（全 19 頁）

(21)出願番号　特願2001－327112(P2001－327112)

(22)出願日　平成13年10月25日(2001.10.25)

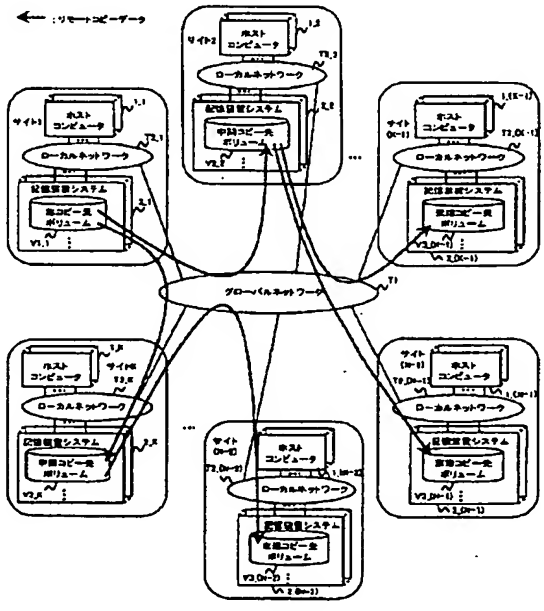
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　森下　昇
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　中野　俊夫
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　記憶装置システム

(57)【要約】
【課題】主に３サイト以上のNサイトに配置した記憶制御装置間のリモートコピーにおいて、任意サイト被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することを課題とする。
【解決手段】ホストコンピュータから直接更新されるリモートコピー範囲と、当該リモートコピー範囲のコピー先となる他サイトのリモートコピー範囲の更新履歴の記録を、当該サイト間で共通に認識した、ホストコンピュータからのある更新を契機に開始し、任意サイトの被災時、被災していない任意サイト間のリモートコピー範囲の内容を、更新履歴に従って、リモートコピー範囲の一部をコピーすることで、リモートコピー範囲の内容を一致させる。

図 1



【特許請求の範囲】
【請求項１】　ホストコンピュータと、ホストコンピュータと接続する記憶装置システムを備える複数のサイトにおいて、サイト間のネットワークにより記憶装置システム同士を接続し、記憶装置システムが、ホストコンピュータから更新されるデータを他サイトの記憶装置システムにも反映するリモートコピー構成において、
ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲と、リモートコピー指定範囲のリモートコピー先となる他サイトの記憶装置システムのコピー先リモートコピー指定範囲の更新履歴の記録を、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共通に認識した、ホストコンピュータからのある更新を契機に開始し、任意サイトの被災時、被災していない任意サイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を、更新履歴に従って、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の一部をコピーすることで、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させることを特徴とする記憶装置システム。
【請求項２】　請求項１に記載の記憶装置システムであって、リモートコピー指定範囲の更新履歴を記録する領域を多重に持ち、
ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲への、ホストコンピュータからの更新量が閾値を超えた契機において、新たな更新履歴への切り替え契機となるホストコンピュータからの更新を決定し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共有し、
各記憶装置システムにおいて、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の更新が、新たな更新履歴への切り替え契機となるホストコンピュータからの更新に到達した時点から、新たな更新履歴の記録を開始し、
リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム全てが、新たな更新履歴の記録を開始したことを確認後、従来の更新履歴の記録を中断し、従来の更新履歴を初期化することを特徴とする、請求項１に記載の記憶装置システム。
【請求項３】　請求項２に記載の記憶装置システムであって、新たな更新履歴への切り替え契機となる、ホストコンピュータからの更新の決定を、前回の更新履歴の切り替えからの経過時間が閾値を超えた契機にも実行することを特徴とする、請求項２に記載の記憶装置システム。
【請求項４】　請求項２に記載の記憶装置システムであって、記憶装置システム間での情報のやりとりを、当該記憶装置システムのリモートコピー指定範囲、または、コピー先リモートコピー指定範囲に対応する直接のコピー元、および、直接のコピー先との間でのみ実行することを特徴とする、請求項２に記載の記憶装置システム。
【請求項５】　請求項２に記載の記憶装置システムであって、記憶装置システム間での情報のやりとりを、リモートコピーデータの転送、および、その応答に添付して実行することを特徴とする、請求項２に記載の記憶装置システム。
【請求項６】　請求項２に記載の記憶装置システムであって、任意サイトの被災時、被災していない任意サイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させるための更新履歴を、
両サイトにおいて、有効、かつ、記録を開始したホストコンピュータからの更新が同じであるという条件で選択し、
さらに、両サイトにおいて、前記条件を充足する複数の更新履歴が存在する場合には、記録を開始したホストコンピュータからの更新が新しい更新履歴を選択することを特徴とする、請求項２に記載の記憶装置システム。
【請求項７】　請求項２に記載の記憶装置システムであって、任意サイトの被災を検出した場合、従来の更新履歴の記録を続行し、被災したサイトがすべて正常となった時点で、当該従来の更新履歴の記録を中断し、初期化することを特徴とする、請求項２に記載の記憶装置システム。
【請求項８】　請求項２に記載の記憶装置システムであって、任意サイトの被災を検出した場合、従来の更新履歴を除外した更新履歴の領域により、被災していないサイト間で請求項２に記載の手順を実行することを特徴とする、請求項２に記載の記憶装置システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ホストコンピュータを介さず、記憶装置システム間でデータを２重化するリモートコピー機能に関する。特に、3サイト以上に配置した記憶装置システム間のリモートコピーにおいて、任意サイトが被災し、被災していないサイト間でリモートコピーを継続する場合における、リモートコピー指定範囲の内容の一致管理に関する。
【０００２】
【従来の技術】災害等による記憶装置システム内のデータ喪失を避けるため、遠隔地にある記憶装置システムに当該データを２重化する機能にリモートコピーがある。
【０００３】リモートコピーでは、ホストコンピュータに接続している正記憶制御装置と、正記憶制御装置と接続している副記憶制御装置に、リモートコピー対象ボリュームを設定し、正記憶制御装置に接続している正記憶装置の正ボリュームと、副記憶制御装置に接続している副記憶装置の副ボリュームの内容を常に一致させるよう

にコピーを実行し続ける。

【０００４】特開平１１－８５４０８号公報には、異なる記憶制御装置間で、ホストコンピュータを介さずデータを2重書きする技術が開示されている。本発明では、ホストコンピュータからライト時刻を付加されたライトデータを受領した正記憶制御装置は、ホストコンピュータにライトデータのライト完了報告後、ライト時刻順に、ライト時刻とライトデータを副記憶制御装置に転送する。副記憶制御装置では、正記憶制御装置から受領したライト時刻とライトデータを不揮発のキャッシュメモリに格納し、あるライト時刻までのデータを保証する。

【０００５】リモートコピーの動作は大きく２つに分けられる。ホストコンピュータへのライト要求の終了報告前に副記憶制御装置にライトデータを転送する同期リモートコピー、ホストコンピュータへのライト要求の終了報告後にライト要求とは非同期に副記憶制御装置にライトデータを転送する非同期リモートコピーである。データベース等のアプリケーションにリモートコピーを使用しようとすると、ログの更新順序を保証するために、正ボリュームに対する更新順序どうりに副ボリュームを更新する必要がある。これを実現するために、非同期リモートコピーでは、正記憶制御装置において、ホストコンピュータからのライト毎に、ライトデータにシーケンス＃を付加し、副記憶制御装置でシーケンス＃どうりに副ボリュームに反映する方法がある。

【０００６】上記に述べた従来のリモートコピーにおいて、リモートコピーを一時中断する場合（例えば、正記憶制御装置と副記憶制御装置間のリンク障害等によってリモートコピーが中断する場合等）、リモートコピー再開時に正ボリュームと副ボリュームの内容を早く一致させることを目的に、リモートコピーが中断している間のホストコンピュータからの正ボリュームに対する更新履歴を正／副ボリューム間の差分情報として正記憶制御装置内に記録しておくことがある。その後、リモートコピー再開時に、記録しておいた更新履歴を使用して、正ボリュームの内容を副ボリュームへ反映することで、正ボリュームと副ボリュームの内容を一致させることができる。

【０００７】従来のリモートコピーにおける差分情報は、通常、ホストコンピュータからの更新位置であって、正ボリュームは更新されているが副ボリュームは更新されていない位置を記録する性質のものであるため、正記憶制御装置側に持つのが一般的である。

【０００８】近年、ネットワークの高速化、低価格化に伴い、遠隔地間のデータ転送コストが下がりつつある。このため、リモートコピーにおける転送データ量を増やし、リモートコピーデータの可用性をさらに高めるために、１サイトが被災してもデータの2重化を維持できるように、３サイト以上のNサイト間でリモートコピーを実施したいという要求がある。

【０００９】Nサイト間リモートコピーの運用として、Nサイトの内１サイトが被災した場合、残りの（N－１）サイトでリモートコピーを再構成し、リモートコピーを継続する形態が望まれる。この場合、リモートコピーを継続するために、残りの（N－１）サイト間でリモートコピーデータを速やかに一致させる必要がある。つまり、（N－１）サイト間の差分管理を行い、差分相当のコピーによって速やかにリモートコピーデータを一致させることが望ましい。

【００１０】なぜなら、差分管理を実施しない場合、全てのリモートコピーデータをコピー元からコピー先へコピーする必要が生じる場合がある。これは、差分相当のコピーと比較し、コピーに長時間かかる上、コピー実行中においてデータの可用性が低下した状態となるため、確保しているサイト数だけの冗長性を生かしきれない。

【００１１】ここで、例として、３サイトでリモートコピーを構成し、従来のリモートコピーにおける差分管理の方法を使用した場合の問題点を考える。３つのサイトをA、B、Cとする。

【００１２】まず、サイト間を転送するリモートコピーデータの流れがA→B→Cの場合、サイトBが被災すると、サイトAとサイトCのボリューム間で差分をコピーしリモートコピーを再開することが望ましい。従来のリモートコピーにおける差分管理情報は正記憶制御装置側に持つのが一般的であることから、サイトA－B間の差分管理情報はサイトAが、サイトB－C間の差分管理情報はサイトBが持つことになる。このため、サイトB被災時には、サイトB－C間の差分管理情報が失われるため、サイトA－C間の差分が分からなくなる。

【００１３】また、サイト間を転送するリモートコピーデータの流れがA→B、A→Cの場合、サイトAが被災すると、サイトBとサイトCのボリューム間で差分をコピーしリモートコピーを再開することが望ましい。この場合、サイトA－B、A－C間の差分管理情報は被災したサイトAが持っているため、サイトB－C間の差分が分からなくなる。

【００１４】この問題に対応するため、通常時にリモートコピーデータのやりとりの無いサイト間で、あらかじめ差分管理を実施しておく方法が考えられる。上記の例においては、サイトＢ被災前にサイトＡ－Ｃ間の差分を、また、サイトＡ被災前にサイトＢ－Ｃ間の差分を管理しておく方法である。しかし、この方法の場合、サイト数に比例してあらかじめ管理しなければならないサイト間の差分情報が増大する、また、例えば、２サイト同時に被災するような想定外の事態に対し差分管理できない。

【００１５】

【発明が解決しようとする課題】従来の、２サイト間でのリモートコピーの差分管理方法は、Nサイト間での差分管理について考慮していない。

【００１６】このため、Nサイト間でのリモートコピー時、1つのサイトが被災し、被災していない残りの（N－１）サイトでリモートコピーを継続しようとした場合、すべてのリモートコピーデータを、新たなコピー元サイトからコピー先サイトへ転送してからでなければ、リモートコピーを再構成できないという問題があった。

【００１７】また、従来の、２サイト間でのリモートコピーの差分管理方法を、サイト被災後のリモートコピー再構成時の新たなサイト間ペアに対してあらかじめ適応しておく場合、サイト数が増加するに従い、各サイトの差分管理情報が増大するという問題があった。また、多重サイトの被災に対して差分管理できないという問題があった。

【００１８】本発明の目的は、主に３サイト以上のNサイトに配置した記憶装置システム間のリモートコピーにおいて、任意サイトの被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することにある。また、リモートコピーを構成するサイト数Nに依存せず、差分を管理する情報量を一定に保つことにある。

【００１９】

【課題を解決するための手段】この目的を達成するために、本発明による記憶装置システムは、ホストコンピュータと、ホストコンピュータと接続する記憶装置システムを備える複数のサイトにおいて、サイト間のネットワークにより記憶装置システム同士を接続し、記憶装置システムが、ホストコンピュータから更新されるデータを他サイトの記憶装置システムにも反映するリモートコピー構成において、ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲と、リモートコピー指定範囲のリモートコピー先となる他サイトの記憶装置システムのコピー先リモートコピー指定範囲の更新履歴を、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共通に認識した、ホストコンピュータからのある更新を契機に記録を開始し、任意サイトの被災時、被災していないサイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を、更新履歴に従って、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の一部をコピーすることで、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させても良い。

【００２０】また、リモートコピー指定範囲の更新履歴を記録する領域を多重に持ち、ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲への、ホストコンピュータからの更新量が閾値を超えた契機において、新たな更新履歴への切り替え契機となるホストコンピュータからの更新を決定し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共有し、各記憶装置システムにおいて、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の更新が、新たな更新履歴への切り替え契機となるホストコンピュータからの更新に到達した時点から、新たな更新履歴の記録を開始し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム全てが、新たな更新履歴の記録を開始したことを確認後、従来の更新履歴の記録を中断し、従来の更新履歴を初期化しても良い。

【００２１】また、新たな更新履歴への切り替え契機となる、ホストコンピュータからの更新の決定を、前回の更新履歴の切り替えからの経過時間が閾値を超えた契機にも実行しても良い。

【００２２】また、記憶装置システム間での情報のやりとりを、当該記憶装置システムのリモートコピー指定範囲、または、コピー先リモートコピー指定範囲に対応する直接のコピー元、および、直接のコピー先との間でのみ実行しても良い。

【００２３】また、記憶装置システム間での情報のやりとりを、リモートコピーデータの転送、および、その応答に添付して実行しても良い。

【００２４】また、任意サイトの被災時、被災していないサイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させるための更新履歴を、両サイトにおいて、有効、かつ、記録を開始したホストコンピュータからの更新が同じであるという条件で選択し、さらに、両サイトにおいて、前記条件を充足する複数の更新履歴が存在する場合には、記録を開始したホストコンピュータからの更新が新しい更新履歴を選択しても良い。

【００２５】また、保守端末より、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲のリモートコピー先毎に、リモートコピー先障害時の他のリモートコピー先へのリモートコピーの実行可否を指定し、リモートコピー先障害時、前記指定に従って、当該記憶装置システムからのリモートコピー先へのリモートコピーを制御しても良い。

【００２６】

【発明の実施の形態】以下、本発明の第1の実施の形態を図面を用いて説明する。

【００２７】図１に、Nサイト間でのリモートコピー構成の示す。各サイトには、複数のホストコンピュータ１や複数の記憶装置システム２等をＬＡＮ（Ｌｏｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）やＳＡＮ（Ｓｔｏｒａｇｅ Ａｒｅａ Ｎｅｔｗｏｒｋ）といったローカルネットワークＴ２によって互いに接続している。一方、各サイトの記憶装置システム２の内、リモートコピーを実行する記憶装置システム２同士はグローバルネットワークＴ１を介し接続している。グローバルネットワークＴ１は、一般的に公衆通信回線であり、通信サービスを提供

する業者から有料で借り受けるものを利用する場合が多い。ただし、ローカルネットワークＴ２、グローバルネットワークＴ１の構成が、本発明を限定するものではない。

【００２８】リモートコピー対象ボリュームの種類は、コピー元ボリュームを持たずコピー先ボリュームを持つ総コピー元ボリュームＶ１、コピー元ボリュームとコピー先ボリュームを持つ中間コピー先ボリュームＶ２、コピー元ボリュームを持ちコピー先ボリュームを持たない末端コピー先ボリュームＶ３に大別できる。リモートコピー対象ボリューム間の関係は、総コピー元ボリュームＶ１を頂点とするツリー状となり、総コピー元ボリュームＶ１から中間コピー先ボリュームＶ２を経て末端コピー先ボリュームＶ３へ順にリモートコピーデータを転送する。

【００２９】図１では、サイト１のボリュームを総コピー元ボリュームＶ１とし、サイト２の中間コピー先ボリュームＶ２を経て、サイト（Ｋ－１）、サイトＮの末端コピー先ボリュームＶ３へ、また、サイトＫの中間コピー先ボリュームＶ２を経て、サイト（Ｎ－１）の末端コピー先ボリュームＶ３へリモートコピーを実施している構成である。

【００３０】図２に、記憶装置システム２の構成を示す。記憶制御装置２１は、ホストコンピュータ１と接続するチャネルインタフェース３と、記憶装置２２と接続するディスクインタフェース４と、管理情報を格納する２重化した不揮発の管理情報メモリ６と、データを格納する２重化した不揮発のキャッシュメモリ５とをパスで接続した構成である。各チャネルインタフェース３、および、各ディスクインタフェース４は、２重化した管理情報メモリ６それぞれ、および、２重化キャッシュメモリ５それぞれと接続している。また、記憶制御装置２１は、当該記憶制御装置２１への指示や当該記憶制御装置２１内部の状態を表示するための保守端末２３を備える。

【００３１】チャネルインタフェース３は、ホストコンピュータ１からのライト要求やリモートコピー元となる他の記憶装置システム２からのリモートコピーのライト要求を処理するライト処理３１、ライトされた範囲に対応するビットマップを更新するビットマップ更新処理３２、自記憶装置システム２が非同期リモートコピーのコピー元である場合、ライトデータをライト要求とは非同期にコピー先である他の記憶装置システム２に転送する非同期転送処理３３、総コピー元ボリュームＶ１を持つ記憶装置システム２から当該ボリュームのコピー先となる全ボリュームのビットマップ８０９を切り替える契機を与えるビットマップ切り替え起動処理３４、ビットマップ切り替え起動処理３４からのビットマップ切り替え要求を受領し、ビットマップを切り替えるビットマップ切り替え処理３５、任意サイトの被災後において残サイト間でリモートコピー構成を再構成する場合の残サイト間のボリューム内容を一致させる差分コピー処理３６を備える。

【００３２】ディスクインタフェース４は、自記憶装置システム２が非同期リモートコピーのコピー先である場合、コピー元からのライト時に仮データとして一旦保存しておいたリモートコピーデータを正式化する非同期正式化処理４１、チャネルインタフェース３側と同じビットマップ更新処理３２を備える。

【００３３】各処理はそれぞれ複数存在し、存在する単位は、ビットマップ切り替え起動処理３４、ビットマップ切り替え処理３５、および、差分コピー処理３６がボリューム単位であったり、非同期転送処理３３、および、非同期正式化処理４１が後述するライトシーケンス管理情報６２単位であったりしても良い。

【００３４】キャッシュメモリ５は、記憶装置２２に格納しているホストコンピュータ１からリード／ライトされるデータを格納している。キャッシュメモリ５の構成は、通常領域と仮領域に大別できる。仮領域を持つ理由は、自記憶装置システム２が非同期リモートコピーのコピー元である場合、コピー先に未転送のリモートコピーデータが再びライトされた際にライト前の未転送データを保存するため、あるいは、自記憶装置システム２が非同期リモートコピーのコピー先である場合、コピー元より受領し非同期正式化処理４１によって正式化するまでのリモートコピーデータを保存するためである。

【００３５】管理情報メモリ６は、記憶装置システム２が動作するために必要な管理情報を格納している。その中には、リモートコピー管理情報６１が含まれ、リモートコピー管理情報６１には、ライトシーケンス管理情報６２、ボリューム管理情報６３が含まれる。

【００３６】ライトシーケンス管理情報６２は、非同期リモートコピーにおいてホストコンピュータ１からのライト順序でリモートコピーデータを正式化するために使用する。ライトシーケンス管理情報６２は複数存在し、各々のライトシーケンス管理情報６２は、ホストコンピュータ１からのライト順序を記録するためのライトシーケンス＃カウンタ７１、ライトデータ毎に割り当てライトシーケンス＃を保存しライトデータに関連付けるライトシーケンス管理情報エントリ領域７２のライトシーケンス管理情報エントリ７２１、非同期リモートコピーのコピー元においてコピー先に未転送のリモートコピーデータを登録する転送対象ライトシーケンス管理情報７３、非同期リモートコピーのコピー先においてコピー元から受領し未だ正式化していないリモートコピーデータを登録する正式化対象ライトシーケンス管理情報７４等から構成する。

【００３７】図３に、ライトシーケンス管理情報６２の構造とキャッシュメモリ５の関係を示す。ライトデータ毎にライトシーケンス管理情報エントリ７２１を割り当

て、ライトシーケンス管理情報エントリ７２１にキャッシュメモリ５のライトデータの位置を設定する。また、転送対象ライトシーケンス管理情報７３、正式化対象ライトシーケンス管理情報７４は、ライトシーケンス管理情報エントリ７２１を接続するキュー構造であっても良い。

【００３８】また、ホストコンピュータ１からのライト順序を複数のボリューム間で保証したい場合には、同じライトシーケンス管理情報６２のライトシーケンス＃カウンタ７１を使用して複数のボリュームのライトデータにライトシーケンス＃を割り当て、同じ転送対象ライトシーケンス管理情報７３に登録し、コピー先において、同じ正式化対象ライトシーケンス管理情報７４に登録し、ライトシーケンス＃に従って正式化すれば良い。

【００３９】図２に戻って、ボリューム管理情報６３は、リモートコピー対象ボリュームに関する管理情報である。ボリューム管理情報６３は複数存在し、各々のボリューム管理情報６３は、自ボリュームが使用するライトシーケンス管理情報６２を指定するライトシーケンス管理情報＃８０１、当該ボリュームのコピー先情報８０２、当該ボリュームのコピー元情報８０３、ビットマップ切り替えを排他するためのビットマップ切り替え中フラグ８０４、当該ボリュームがコピー先ボリュームである場合の最新ライトシーケンス＃８０５、最新ライトシーケンス＃と比較しビットマップを切り替える契機とするビットマップ切り替え契機ライトシーケンス＃８０６、ビットマップ切り替え時に自記憶装置システム２において最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６に達したかどうかの確認を要求するライトシーケンス＃確認要求フラグ８０７、最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６に達したことを示すライトシーケンス＃到達フラグ８０８、当該ボリュームの更新箇所を記録するビットマップ８０９（複数）、対応するビットマップが有効であることを示すビットマップ有効フラグ８１０（複数）、対応するビットマップへの記録を開始したライトシーケンス＃を示すビットマップ記録開始ライトシーケンス＃８１１（複数）、対応するビットマップへの更新量を示すビットマップ更新データ量カウンタ８１２（複数）、サイト被災等により他サイトとのボリューム内容を一致させる必要が生じた場合に使用する差分ビットマップ８１３等から構成する。

【００４０】コピー先情報８０２、コピー元情報８０３とは、コピー元／先の記憶装置システム２と通信するために必要な情報や、コピー元／先のボリュームを特定する情報等である。

【００４１】また、簡単のため、本発明の実施の形態においては、管理情報をボリューム単位に持つ（ボリューム管理情報６３）こととしたが、これは、本発明の範囲を制限するものではない。管理情報は、記憶装置システム２間で共有できる任意の単位で持つことができ、よって、記憶装置システム２間で共有できる任意の単位でリモートコピーを実行することができる。図４に、各記憶装置システム２で実行するライト処理３１について説明する。

【００４２】説明を分かり易くするため、ライト処理３１が対象とする、ライト要求元とライトデータの種別の概要を補足する。ライト要求元は、ホストコンピュータ１、または、リモートコピー元の記憶装置システム２である。

【００４３】ホストコンピュータ１からのライトデータは、ライト対象ボリュームにより、１）リモートコピー先の無いもの、リモートコピー先の有るもので２）同期リモートコピー先のみのもの、３）非同期リモートコピー先のみのもの、４）両方のものに分けられる。

【００４４】リモートコピー元の記憶装置システム２からのライトデータは、ライト対象ボリュームにより、同期リモートコピーによりライトされたもの、非同期リモートコピーによりライトされたものに分けられ、同期リモートコピーによりライトされたものは、上記１）～４）と同様に、５）リモートコピー先の無いもの、リモートコピー先の有るもので６）同期リモートコピー先のみのもの、７）非同期リモートコピー先のみのもの、８）両方のものに分けられる。また、非同期リモートコピーによるものは、上記１）、４）と同様に、９）リモートコピー先の無いもの、１０）非同期リモートコピー先のみのものに分けられる。２）、３）を対象外としたのは、同期リモートコピーがリモートコピー完了後にホストコンピュータにライト終了を報告するものであり、中間コピー先ボリュームＶ２において、非同期リモートコピーで受領したリモートコピーデータを同期リモートコピーで他サイトのボリュームにリモートコピーすることは、あまり意味が無く現実的でないためである。

【００４５】ライト処理３１では、まず、ライト要求を受領し（ステップＳ４０１）、ライト要求元がホストコンピュータ１、かつ、ライト対象ボリュームがリモートコピー元でないか確認する（ステップＳ４０２）。

【００４６】ステップＳ４０２の条件成立の場合、ホストコンピュータ１からリモートコピーに関連しないボリュームへのライト要求であるため（上記１）に相当）、ライトデータをキャッシュメモリ５の通常領域に格納し（ステップＳ４０３）、ライト要求元にライト終了を報告する（ステップＳ４０４）。このライトデータは、ホストコンピュータ１からのライト要求とは非同期に、記憶制御装置２１が記憶装置２２に書き込む。

【００４７】ステップＳ４０２の条件不成立の場合、ライト要求元がホストコンピュータ１であるか確認する（ステップＳ４０５）。

【００４８】ステップＳ４０５の条件成立の場合（上記２）３）４）に相当）、ライトデータに対応させるライ

トシーケンス管理情報エントリ７２１を確保し、ライトシーケンス＃を付与する。具体的には、ホストコンピュータ１からライトされたボリュームに対応するボリューム管理情報６３のライトシーケンス管理情報＃８０１を参照し、対応するライトシーケンス管理情報６２のライトシーケンス管理情報エントリ７２１を確保し、ライトシーケンス＃カウンタ７１の値を設定し、ライトシーケンス＃カウンタ７１の値を加算し、ライトシーケンス管理情報エントリ７２１をライトデータと関連付ける。

【００４９】ライトシーケンス＃の付加は、リモートコピーの同期／非同期に関係なく実行する。これは、仮に当該ボリュームの直接のコピー先が同期リモートコピー先だけであったとしても、これらのコピー先のさらに先が非同期リモートコピーを実行することがあるためである。

【００５０】ステップＳ４０５の条件不成立の場合（上記6）～１０）に相当）、ライトデータに対応させるライトシーケンス管理情報エントリ７２１を確保し、ライトデータと共に受領したライトシーケンス＃を含むライトシーケンス管理情報６２を格納する（ステップＳ４０７）。

【００５１】次に、ライトデータをキャッシュメモリ５に格納する。キャッシュメモリ５への格納方法は、ライト対象ボリュームにより異なり、A）非同期リモートコピー先である（上記９）１０）に相当）、B）非同期リモートコピー先ではなく、非同期リモートコピー元である（上記３）４）７）８）に相当）、C）それら以外（上記２）５）６）に相当）の３通りに分けられる。これは、ライト対象ボリュームがA）である場合、受領したライトデータを一旦仮保存し、ライトシーケンス＃に従って正式にライトする必要があるためである。また、B）である場合、非同期リモートコピー先に未転送のライトデータをライトされた場合、古いライトデータを転送するまで別に保存しておく必要があるためである。

【００５２】まず、ライト対象ボリュームが非同期リモートコピー先かどうか判断する（ステップＳ４０８）。

【００５３】ステップＳ４０８の条件成立（上記A）に相当）の場合、当該ライトデータはホストコンピュータ１からの更新順を示すライトシーケンス＃に従ってライトする必要があるため、一旦、ライトデータをキャッシュメモリ５の仮領域に格納し（ステップＳ４０９）、ライトデータに対応するライトシーケンス管理情報エントリ７２１を正式化対象データとして正式化対象ライトシーケンス管理情報７４に登録する（ステップＳ４１０）。図３においては、ライトシーケンス管理情報エントリ７２１＿ｅ、７２１＿ｆが相当する。

【００５４】さらに、ライト対象ボリュームが非同期リモートコピーのコピー元であるか判断し（ステップＳ４１１）、ステップＳ４１１の条件成立（上記３）４）７）８）１０）に相当）の場合、ライトデータに対応するライトシーケンス管理情報エントリ７２１を転送対象データとして転送対象ライトシーケンス管理情報７３に登録する（ステップＳ４１２）。図３においては、ライトシーケンス管理情報エントリ７２１＿ｂ、７２１＿ｃ、７２１＿ｄ、７２１＿ｅが相当する。

【００５５】ステップＳ４０８の条件成立の場合、さらに、ライト対象ボリュームが非同期リモートコピーのコピー元であるか判断する（ステップＳ４１３）。

【００５６】ステップＳ４１３の条件成立（上記B）に相当）の場合、非同期リモートコピー先へ未転送のライトデータをキャッシュメモリ５の仮領域へ移設した上で、ライトデータをキャッシュメモリ５の通常領域に格納する（ステップＳ４１４）。図３においては、ライトシーケンス管理情報エントリ７２１＿ｂ、７２１＿ｃ、７２１＿ｄが相当する。さらに、ビットマップ更新処理３２を実行する（ステップＳ４１５）。ビットマップ更新処理３２については後述する。その後、ステップＳ４１１へ。

【００５７】ステップＳ４１３の条件不成立（上記C）に相当）の場合、当該ライトデータは、非同期リモートコピーのコピー先でもコピー元でもないため、ライトデータをキャッシュメモリ５の通常領域に格納する（ステップＳ４１６）。図３においては、ライトシーケンス管理情報エントリ７２１＿ａが相当する。さらに、ビットマップ更新処理３２を実行する（ステップＳ４１７）。ビットマップ更新処理３２については後述する。その後、ステップＳ４１８へ。

【００５８】次に、ライト対象ボリュームが同期リモートコピーのコピー元であるか判断し（ステップＳ４１８）、ステップＳ４１８の条件成立の場合（上記２）４）６）８）に相当）、同期リモートコピーのコピー先にライトデータとライトシーケンス管理情報エントリ７２１の内容を転送する（ステップＳ４１９）。当該同期リモートコピーは、ライト要求元へのライト要求を受領してからライト終了を報告するまでの時間を短縮するために、複数のコピー先に対し並列に実行しても良い。

【００５９】次に、当該ライトデータに対応するライトシーケンス管理情報７２１が、転送対象ライトシーケンス管理情報７３、および、正式化対象ライトシーケンス管理情報７４のいずれにも登録されていないか判断する（ステップＳ４２０）。ステップＳ４２０の条件成立（上記２）５）６）に相当）の場合、当該ライトデータに対応するライトシーケンス管理情報エントリ７２１は不要であるため、ライトシーケンス管理情報エントリ７２１を解放する（ステップＳ４２１）。図３においては、ライトシーケンス管理情報エントリ７２１＿ｇが相当する。

【００６０】最後に、ライト要求元にライト終了を報告（ステップＳ４０４）し、処理を終了する。

【００６１】図５に、記憶装置システム２の内、非同期

リモートコピーのコピー元ボリュームを持つ記憶装置システム２で実行する非同期転送処理３３について説明する。当該処理は、ライトシーケンス管理情報６２単位に実行しても良い。

【００６２】まず、転送対象ライトシーケンス管理情報７３に登録されたライトシーケンス管理情報エントリ７２１から転送対象を選択する（ステップＳ５０１）。

【００６３】次に、非同期リモートコピー先に、選択したライトシーケンス管理情報エントリ７２１と対応するライトデータを転送する（ステップＳ５０２）。

【００６４】次に、転送したライトシーケンス管理情報エントリ７２１の転送対象ライトシーケンス管理情報７３への登録を解除する（ステップＳ５０３）。

【００６５】次に、転送したライトシーケンス管理情報エントリ７２１が正式化対象ライトシーケンス管理情報７４にも登録されているか判断する（ステップＳ５０４）。ステップＳ５０４の条件不成立の場合、非同期リモートコピー対象データとしての管理が不要となるため、転送したライトシーケンス管理情報エントリ７２１を解放し（ステップＳ５０５）、転送したライトデータがキャッシュメモリ５の仮領域を使用していた場合には、当該仮領域も解放する（ステップＳ５０６）。

【００６６】図６に、記憶装置システム２の内、非同期リモートコピーのコピー先ボリュームを持つ記憶装置システム２で実行する非同期正式化処理４１について説明する。当該処理は、ライトシーケンス管理情報６２単位に実行しても良い。

【００６７】まず、正式化対象ライトシーケンス管理情報７４に登録されたライトシーケンス管理情報エントリ７２１を参照し、ライトシーケンス＃が連続している範囲を正式化対象として選択する（ステップＳ６０１）。

【００６８】次に、選択した正式化対象のライトデータは全て正式化済みか判断する（ステップＳ６０２）。

【００６９】ステップＳ６０２の条件成立の場合、ステップＳ６０６へ。

【００７０】ステップＳ６０２の条件不成立の場合、正式化対象として選択したライトデータの内、最も古いライトシーケンス＃に対応するライトデータをキャッシュメモリ５の仮領域から通常領域に移設し、正式化する（ステップＳ６０３）。

【００７１】次に、正式化したライトシーケンス管理情報エントリ７２１の正式化対象ライトシーケンス管理情報７４への登録を解除する（ステップＳ６０４）
次に、ビットマップ更新処理３２を実行する（ステップＳ６０５）。ビットマップ更新処理３２については後述する。その後、ステップＳ６０２へ戻る。

【００７２】次に、正式化したライトシーケンス管理情報エントリ７２１が転送対象ライトシーケンス管理情報７３にも登録されているか判断する（ステップＳ６０６）。ステップＳ６０６の条件不成立の場合、非同期リモートコピー対象データとしての管理が不要となるため、正式化したライトシーケンス管理情報エントリ７２１を解放し（ステップＳ６０７）、正式化したライトデータが使用していたキャッシュメモリ５の仮領域を解放する（ステップＳ６０８）。

【００７３】図７に、記憶装置システム２の内、総コピー元ボリュームＶ１を持つ記憶装置システム２で実行するビットマップ切り替え起動処理３４について説明する。当該処理は、ビットマップ有効フラグ８１０がＯＮであるビットマップ８０９（２つ以上存在する場合には記録開始ライトシーケンス＃８１１の古い方）の更新データ量カウンタ８１２の値が一定の閾値を超えた場合に、ボリューム単位に実行しても良い。

【００７４】また、前回実行したビットマップ切り替えからの経過時間が一定の閾値を超えた場合に、ボリューム単位に実行しても良い。

【００７５】まず、総コピー元ボリュームＶ１、および、当該ボリュームを総コピー元ボリュームＶ１とする全ての中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３に対して、前回のビットマップ切り替え終了前に新たにビットマップ切り替えを開始しないようにするため、ビットマップ切り替え中フラグ８０４がＯＦＦであるか判断する（ステップＳ７０１）。

【００７６】ステップＳ７０１の条件不成立の場合、未だビットマップ切り替え中のため、処理を終了する。

【００７７】ステップＳ７０１の条件成立の場合、ビットマップ切り替え中フラグ８０４をＯＮする（ステップＳ７０２）。

【００７８】次に、ビットマップ切り替え契機とするライトシーケンス＃を決定し、管理情報メモリのビットマップ切り替え契機ライトシーケンス＃８０６に設定する（ステップＳ７０３）。ビットマップ切り替え契機とするライトシーケンス＃の決定方法として、総コピー元ボリュームＶ１から、当該ボリュームを総コピー元ボリュームＶ１とする全ての中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３へ通知するまでに、中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３の最新ライトシーケンス＃８０５が超えてしまわない程度に進んだライトシーケンス＃を生成し、ビットマップ切り替え契機ライトシーケンス＃８０６として良い。

【００７９】また、ビットマップ切り替え契機ライトシーケンス＃８０６が、切り替え先ビットマップ８０９の記録開始ライトシーケンス＃となることを意図しているが、１つのライトシーケンス管理情報６２で複数のボリュームのライトシーケンスを管理する場合、決定したビットマップ切り替え契機ライトシーケンス＃８０６が必ずしも当該ボリュームのライトデータに対して割り当てられるとは限らない。しかし、決定したビットマップ切り替え契機ライトシーケンス＃８０６を超えて当該ボリュームに最初に割り当てられるライトシーケンス＃は、

総コピー元ボリュームＶ１と、当該ボリュームを総コピー元ボリュームＶ１とする全ての中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３において、一意に定まるため、ビットマップ切り替え契機ライトシーケンス＃８０６を超えて当該ボリュームに最初に割り当てられるライトシーケンス＃を、切り替え先ビットマップ８０９の記録開始ライトシーケンス＃とすることができる。

【００８０】次に、最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６と同じとなった、あるいは、超えたことを知らせるライトシーケンス＃到達フラグ８０８をＯＦＦする（ステップＳ７０４）。

【００８１】次に、ビットマップ更新処理３２に最新ライトシーケンス＃８０５とビットマップ切り替え契機ライトシーケンス＃８０６の比較を要求するライトシーケンス＃確認要求フラグ８０７をＯＮする（ステップＳ７０５）。

【００８２】次に、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替え契機ライトシーケンス＃８０６を転送し、最新ライトシーケンス＃がビットマップ切り替え契機ライトシーケンス＃８０６を超えた場合のライトシーケンス＃到達報告を要求する（ステップＳ７０６）。

【００８３】次に、当該ボリュームのライトシーケンス＃到達フラグ８０８がＯＮ、かつ、当該ボリュームの直接のコピー先全てからライトシーケンス＃到達報告が有るか判断する（ステップＳ７０７）。ステップＳ７０７の条件不成立の場合、一定時間待ち（ステップＳ７０８）、再びステップＳ７０７の判断を実行する。なお、当該ボリュームの直接のコピー先からの報告は、逐次、コピー先情報８０２に記録するものとする。

【００８４】ステップＳ７０７の条件成立の場合、自サイトのビットマップ８０９の内、記録開始ライトシーケンス＃８１１が古い方の有効フラグ８１０をＯＦＦし、当該ビットマップ８０９をクリアする（ステップＳ７０９）。これは、この時点で、当該ボリュームと、当該ボリュームを総コピー元ボリュームＶ１とする全ての中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３の最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６を超え、新たなビットマップ８０９への記録を開始済みであるためである。

【００８５】次に、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替えを要求する（ステップＳ７１０）。

【００８６】次に、当該ボリュームの直接のコピー先全てからビットマップ切り替え完了報告あるか判断する（ステップＳ７１１）。ステップＳ７１１の条件不成立の場合、一定時間待ち（ステップＳ７１２）、再びステップＳ７１１の判断を実行する。

【００８７】ステップＳ７１１の条件成立の場合、ビットマップ切り替え終了のため、ビットマップ切り替え中フラグ８０４をＯＦＦし（ステップＳ７１３）、処理を終了する。

【００８８】図８に、記憶装置システム２の内、中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３を持つ記憶装置システム２で実行するビットマップ切り替え処理３５について説明する。

【００８９】まず、コピー元サイトからビットマップ切り替え契機ライトシーケンス＃８０６を受領し、ビットマップ切り替え対象となる中間コピー先ボリュームＶ２と末端コピー先ボリュームＶ３のビットマップ切り替え契機ライトシーケンス＃８０６に設定する（ステップＳ８０１）。

【００９０】次に、ライトシーケンス＃到達フラグ８０８をＯＦＦし（ステップＳ８０２）、ライトシーケンス＃確認要求フラグ８０７をＯＮする（ステップＳ８０３）。

【００９１】次に、当該ボリュームに直接のコピー先が存在するか判断する（ステップＳ８０４）。

【００９２】ステップＳ８０４の条件不成立（末端コピー先ボリュームＶ３に相当）の場合、ステップＳ８０８へ。

【００９３】ステップＳ８０４の条件成立（中間コピー先ボリュームＶ２に相当）の場合、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替え契機ライトシーケンス＃８０６を転送し、最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６を超えた場合のライトシーケンス＃到達報告を要求する（ステップＳ８０５）。

【００９４】次に、当該ボリュームの直接のコピー先全てからライトシーケンス＃到達報告が有るか判断する（ステップＳ８０６）。ステップＳ８０６の条件不成立の場合、一定時間待ち（ステップＳ８０７）、再びステップＳ８０６の判断を実行する。

【００９５】ステップＳ８０６の条件成立の場合、当該ボリュームのライトシーケンス＃到達フラグがＯＮであるか判断する（ステップＳ８０８）。ステップＳ８０８の条件不成立の場合、一定時間待ち（ステップＳ８０９）、再びステップＳ８０８の判断を実行する。

【００９６】ステップＳ８０８の条件成立の場合、当該ボリュームの直接のコピー元にライトシーケンス#到達を報告する（ステップＳ８１０）。

【００９７】次に、当該ボリュームの直接のコピー元からビットマップ切り替え要求を受領したか判断する（ステップＳ８１１）。ステップＳ８１１の条件不成立の場合、一定時間待ち（ステップＳ８１２）、再びステップＳ８１１の判断を実行する。

【００９８】ステップＳ８１１の条件成立の場合、当該ボリュームのビットマップ８０９の内、記録開始ライト

シーケンス＃８１１が古い方のビットマップ有効フラグ８１０をＯＦＦし、当該ビットマップ８０９をクリアする（ステップＳ８１３）。

【００９９】次に、当該ボリュームに直接のコピー先が存在するか判断する（ステップＳ８１４）。

【０１００】ステップＳ８１４の条件不成立（末端コピー先ボリュームＶ３に相当）の場合、ステップＳ８１８へ。

【０１０１】ステップＳ８１４の条件成立（中間コピー先ボリュームＶ２に相当）の場合、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替えを要求する（ステップＳ８１５）。

【０１０２】次に、当該ボリュームの直接のコピー先全てからビットマップ切り替え完了報告が有るか判断する（ステップＳ８１６）。ステップＳ８１６の条件不成立の場合、一定時間待ち（ステップＳ８１７）、再びステップＳ８１６の判断を実行する。

【０１０３】ステップＳ８１６の条件成立の場合、当該ボリュームの直接のコピー元にビットマップ切り替え完了を報告し（ステップＳ８１８）、処理を終了する。

【０１０４】図９に、ライト処理３１、非同期正式化処理４１で実行するビットマップ更新処理３２について説明する。

【０１０５】まず、ビットマップ有効フラグ８１０がＯＮであるビットマップ８０９のライトデータに対応するビットをＯＮし、更新したビットマップ８０９に対応した更新データ量カウンタ８１２をインクリメントする（ステップＳ９０１）。

【０１０６】次に、最新ライトシーケンス＃８０５を当該ライトデータに対応したライトシーケンス＃に更新する（ステップＳ９０２）。

【０１０７】次に、ライトシーケンス＃確認要求フラグ８０７がＯＮであるか判断する（ステップＳ９０３）。

【０１０８】ステップＳ９０３の条件不成立の場合、処理を終了する。

【０１０９】ステップＳ９０３の条件成立の場合、最新ライトシーケンス＃８０５がビットマップ切り替え契機ライトシーケンス＃８０６と同じ、あるいは、超えるか判断する（ステップＳ９０４）。

【０１１０】ステップＳ９０４の条件不成立の場合、処理を終了する。

【０１１１】ステップＳ９０４の条件成立の場合、ライトシーケンス＃到達フラグ８０８をＯＮする（ステップＳ９０５）。

【０１１２】次に、ビットマップ有効フラグ８１０がＯＦＦであるビットマップ８０９のビットマップ有効フラグ８１０をＯＮし、記録開始ライトシーケンス＃８１１に最新ライトシーケンス＃８０５を設定する（ステップＳ９０６）。

【０１１３】次に、新たにビットマップ有効フラグ８１０をＯＮしたビットマップ８０９におけるライトデータに対応するビットをＯＮし、更新したビットマップ８０９に対応した更新データ量カウンタ８１２をインクリメントする（ステップＳ９０７）。

【０１１４】最後に、ライトシーケンス＃確認要求フラグ８０７をＯＦＦし（ステップＳ９０８）、処理を終了する。

【０１１５】図１０に、ビットマップ切り替えにおける、各サイトの処理とサイト間の通信の概要を示す。

【０１１６】総コピー元ボリュームＶ１を持つサイトで実行しているのがビットマップ切り替え起動処理３４、中間コピー先ボリュームＶ２、または、末端コピー先ボリュームＶ３を持つサイトで実行しているのがビットマップ切り替え処理３５である。ただし、一部ビットマップ更新処理３２を含む。各実行内容には、該当するステップ番号を付加した。

【０１１７】図１０は、サイト全体のビットマップ１からビットマップ２へのビットマップ切り替えの概要を示しており、ビットマップ切り替えは、この後、ビットマップ２からビットマップ１へ、さらに、ビットマップ１からビットマップ２へといったように、交互に実行する。

【０１１８】ビットマップ切り替えを、サイト全体として、単一のビットマップを更新する状態から、一時的に２重のビットマップを更新する状態を経て、再び単一のビットマップを更新する状態へ移行することにより、ビットマップ切り替え中を含む任意の瞬間、任意のサイト同士で必ず記録開始ライトシーケンス＃８１１が同じビットマップ８０９を共有することができ、ライトシーケンス＃の進んだサイト側にボリューム内容を一致させることができる。

【０１１９】また、ビットマップ切り替えにおいて、一部のサイトの障害を検出した場合、ビットマップ切り替え契機ライトシーケンス＃８０６の各サイトの記憶装置システム２への転送が失敗することがある。この場合、総コピー元ボリュームＶ１を持つ記憶装置システム２において、ステップＳ７０９のビットマップ８０９クリアの内容を実行せず、ステップＳ７１０において、ビットマップ８０９をクリアせずビットマップ切り替え完了を報告するよう他サイトの記憶装置システム２に要求することで、障害を発生したサイトも含めた差分の管理を続行することができる。

【０１２０】さらに、各記憶装置システム２においてビットマップ８０９を３重以上持ち、一部サイトの障害を検出した場合、上記のように、障害を検出した時点で使用していたビットマップ８０９をクリアしないようにし、残りのビットマップを使用してビットマップ切り替えを続行することで、障害を発生したサイトも含めた差分の管理を続行しながら、正常なサイト間の差分の増大を抑止することができる。クリアしないようにしたビッ

トマップ８０９を、すべてのサイトが正常となるまで維持することで、障害サイトがさらに増えても、障害を発生したサイトも含めた差分の管理を続行することができる。

【０１２１】また、ビットマップ切り替えに伴う記憶装置システム２間の通信内容を、通常のリモートコピーデータの転送、および、応答に添付して実行することで、サイト間の通信負荷を削減できる。

【０１２２】図１１に、サイト被災後、サイト間でボリューム内容を一致させるための手順について説明する。

【０１２３】まず、サイト被災により直接のコピー元を失った自ボリュームの新たな直接のコピー元として、自ボリュームと同じ総コピー元ボリュームＶ１を持つボリュームの内、自ボリュームより最新ライトシーケンス＃８０５が進んでいるボリュームを選択する（ステップＳ１１０１）。

【０１２４】次に、自サイトと新たな直接のコピー元サイトの記憶装置システム２間に転送パスを設定する（ステップＳ１１０２）。

【０１２５】次に、自サイトと新たな直接のコピー元サイトの記憶装置システム２間でリモートコピーを開始する（ステップＳ１１０３）。これ以降、新たな直接のコピー元サイトのコピー元ボリュームに対してなされた更新は、自サイトのコピー先ボリュームに逐次反映される。

【０１２６】最後に、自サイトと新たな直接のコピー元サイトの記憶装置システム２間で差分コピー処理３６を起動する（ステップＳ１１０４）。差分コピー処理３６については後述する。

【０１２７】差分コピー処理３６終了後も、そのままリモートコピーを継続することで、通常と同じリモートコピー動作となる。

【０１２８】上記の新たな直接のコピー元サイトとして、あらかじめ、自ボリュームについて自サイトより総コピー元ボリュームＶ１に近い上流側のサイトを登録しておいても良い。

【０１２９】図１２に、サイト被災後、サイト間でボリューム内容を一致させるための手順の1つである差分コピー処理３６について説明する。

【０１３０】まず、新たな直接のコピー元サイトにおいて、自サイトと新たな直接コピー元サイトの両サイトでビットマップ有効フラグ８１０がＯＮであるビットマップ８０９で、かつ、この条件を満たすビットマップが複数存在する場合には、記録開始ライトシーケンス＃８１１が新しい方のビットマップ８０９を選択し、ビットマップ８０９の内容を、差分コピー処理３６でコピーするデータを決定するための差分ビットマップ８１３にコピーする（ステップＳ１２０１）。差分ビットマップ８１３にコピーする理由は、新たな直接のコピー元サイトでは、サイト被災後もリモートコピーが継続している場合があり、ビットマップ有効フラグ８１０がＯＮであるビットマップ８０９が更新され続けることがあるため、差分コピー量の増大を防ぐため、現状の両サイトのボリューム間の差分を保存しておく必要があるためである。

【０１３１】次に、差分ビットマップ８１３でＯＮとなっているビットに対応する部分を、新たな直接のコピー元サイトのコピー元ボリュームから自サイトのコピー先ボリュームにコピーする（ステップＳ１２０２）。

【０１３２】最後に、差分ビットマップ８１３をクリアし（ステップＳ１２０３）、処理を終了する。

【０１３３】

【発明の効果】本発明によれば、主に３サイト以上のNサイトに配置した記憶制御装置間のリモートコピーにおいて、１サイト被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することができる。また、リモートコピーを構成するサイト数Nに依存せず、差分を管理する情報量を一定に保つことができる。

【図面の簡単な説明】

【図１】本発明の実施の形態における、Ｎサイト間のリモートコピーの構成例を示す。

【図２】本発明の実施の形態における、記憶装置システムの構成例を示す。

【図３】本発明の実施の形態における、ライトシーケンス管理情報とキャッシュメモリの関係の一例を示す。

【図４】本発明の実施の形態における、記憶装置システムで実行するライト処理の一例を示す。

【図５】本発明の実施の形態における、非同期リモートコピー元の記憶装置システムで実行する非同期転送処理の一例を示す。

【図６】本発明の実施の形態における、非同期リモートコピー先の記憶装置システムで実行する非同期正式化処理の一例を示す。

【図７】本発明の実施の形態における、総コピー元ボリュームＶ１を持つ記憶装置システムで実行するビットマップ切り替え起動処理の一例を示す。

【図８】本発明の実施の形態における、中間／末端コピー元ボリュームを持つ記憶装置システムで実行するビットマップ切り替え処理の一例を示す。

【図９】本発明の実施の形態における、ライト処理、および、非同期正式化処理で実行するビットマップ更新処理の一例を示す。

【図１０】本発明の実施の形態における、ビットマップ切り替えの概要を示す。

【図１１】本発明の実施の形態における、サイト被災後、サイト間でボリューム内容を一致させるための手順の一例を示す。

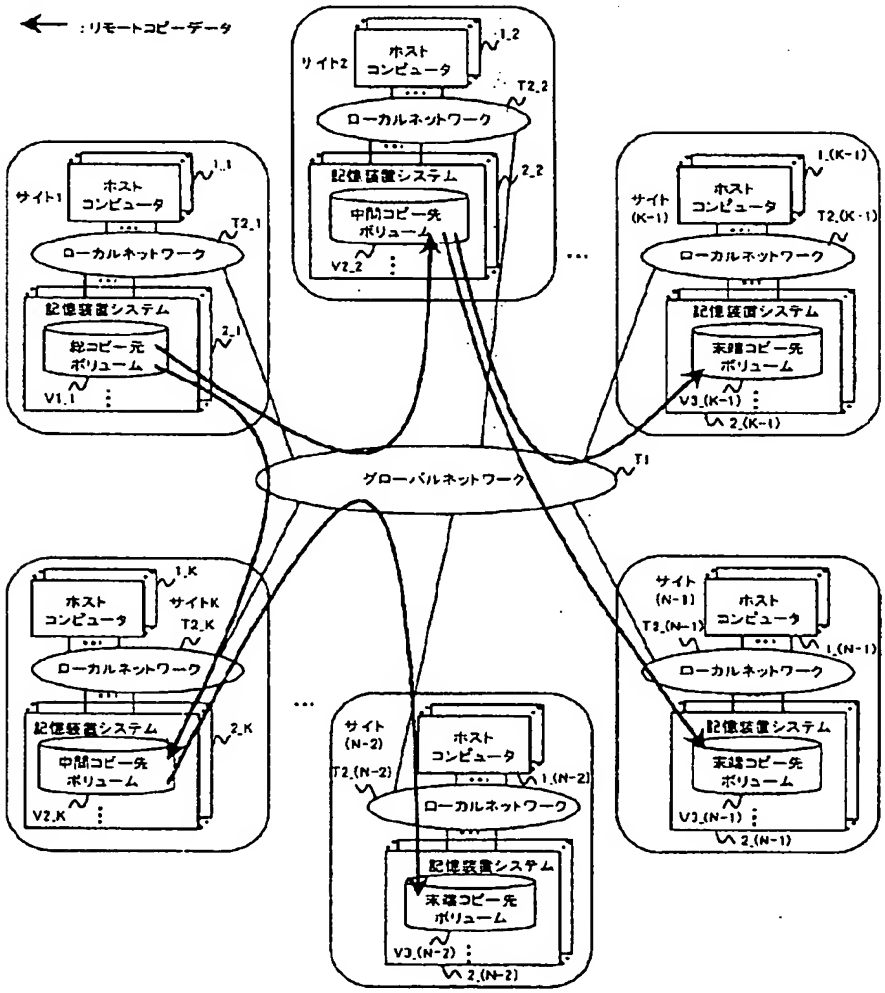
【図１２】本発明の実施の形態における、サイト被災後、サイト間でボリューム内容を一致させるための手順の1つである差分コピー処理の一例を示す。

Japanese Patent Laid-Open No. 2003-131917
English Translation of the wordings in Figs. 1-12 are attached.

【符号の説明】
１…ホストコンピュータ、２…記憶装置システム、２１…記憶制御装置、２２…記憶装置、２３…保守端末、３…チャネルインタフェース、３１…ライト処理、３２…ビットマップ更新処理、３３…非同期転送処理、３４…ビットマップ切り替え起動処理、３５…ビットマップ切り替え処理、３６…差分コピー処理、４…ディスクインタフェース、４１…非同期正式化処理、５…キャッシュメモリ、６…管理情報メモリ、６２…ライトシーケンス管理情報、６３…ボリューム管理情報、７１…ライトシーケンス＃カウンタ、７２１…ライトシーケンス管理情報エントリ、７３…転送対象ライトシーケンス管理情報、７４…正式化対象ライトシーケンス管理情報、８０４…ビットマップ切り替え中フラグ、８０５…最新ライトシーケンス＃、８０６…ビットマップ切り替え契機ライトシーケンス＃、８０７…ライトシーケンス＃確認要求フラグ、８０８…ライトシーケンス＃到達フラグ、８０９…ビットマップ、８１０…ビットマップ有効フラグ、８１１…ビットマップ記録開始ライトシーケンス＃、８１２…ビットマップ更新量データカウンタ、８１３…差分ビットマップ、Ｔ１…グローバルネットワーク、Ｔ２…ローカルネットワーク、Ｖ１…総コピー元ボリューム、Ｖ２…中間コピー先ボリューム、Ｖ３…末端コピー先ボリューム。

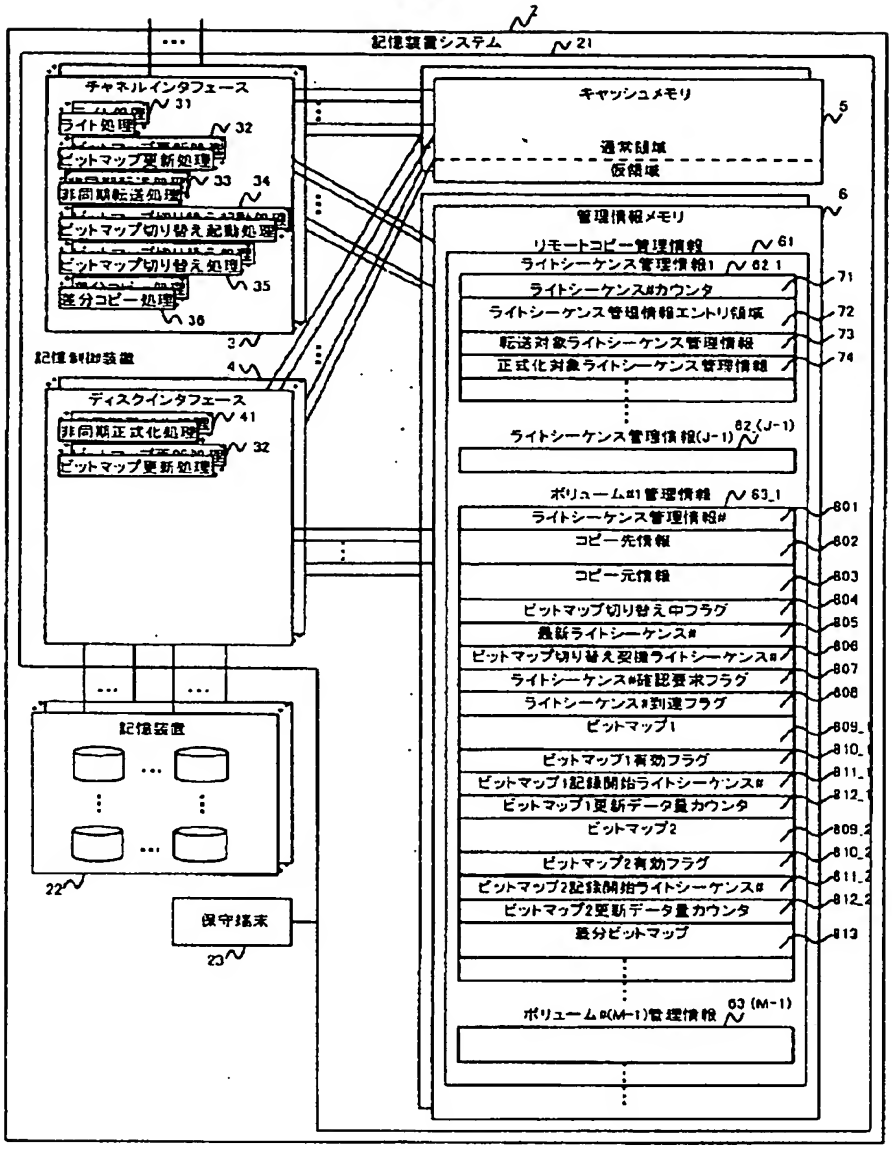
【図１】

図　1



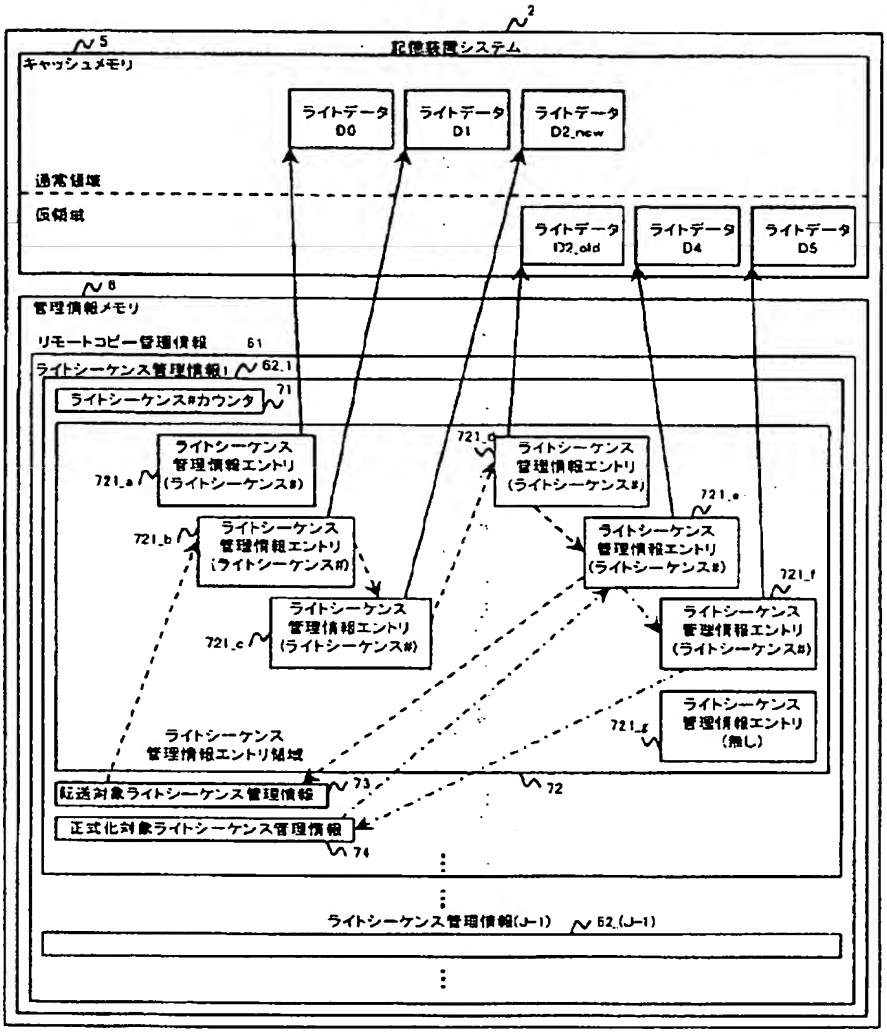
【図２】

図　２


記憶装置システム
21
チャネルインタフェース
31
ライト処理
32
ビットマップ更新処理
33
34
非同期転送処理
ビットマップ切り替え起動処理
ビットマップ切り替え処理
35
差分コピー処理
36
3
記憶制御装置
4
ディスクインタフェース
41
非同期正式化処理
32
ビットマップ更新処理
記憶装置
22
保守端末
23
キャッシュメモリ
5
通常領域
仮領域
6
管理情報メモリ
リモートコピー管理情報
61
ライトシーケンス管理情報1
62_1
71
ライトシーケンス#カウンタ
ライトシーケンス管理情報エントリ領域
72
転送対象ライトシーケンス管理情報
73
正式化対象ライトシーケンス管理情報
74
ライトシーケンス管理情報(J-1)
62_(J-1)
ボリューム#1管理情報
63_1
801
ライトシーケンス管理情報#
802
コピー先情報
803
コピー元情報
804
ビットマップ切り替え中フラグ
805
最新ライトシーケンス#
806
ビットマップ切り替え契機ライトシーケンス#
807
ライトシーケンス#確認要求フラグ
808
ライトシーケンス#到達フラグ
809_1
ビットマップ1
810_1
ビットマップ1有効フラグ
811_1
ビットマップ1記録開始ライトシーケンス#
812_1
ビットマップ1更新データ量カウンタ
809_2
ビットマップ2
810_2
ビットマップ2有効フラグ
811_2
ビットマップ2記録開始ライトシーケンス#
812_2
ビットマップ2更新データ量カウンタ
813
差分ビットマップ
ボリューム#(M-1)管理情報
63_(M-1)

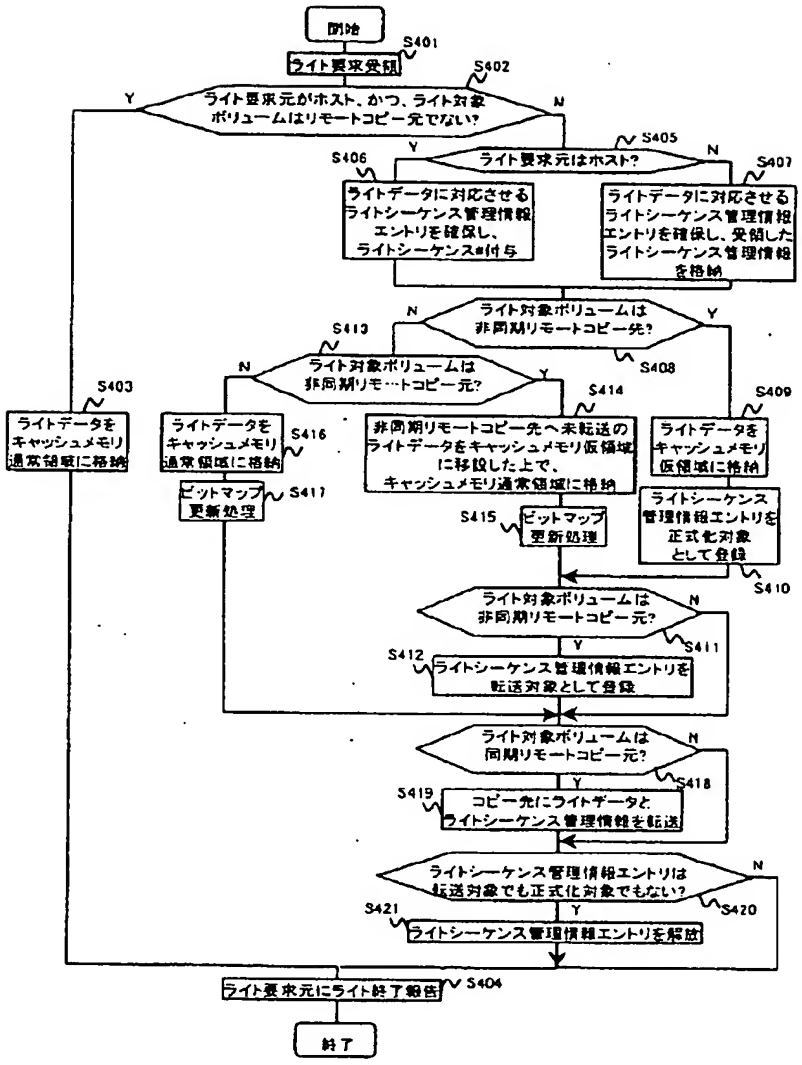
【図３】

図　３


2
記憶装置システム
5
キャッシュメモリ
ライトデータ D0
ライトデータ D1
ライトデータ D2_new
通常領域
仮領域
ライトデータ D2_old
ライトデータ D4
ライトデータ D5
6
管理情報メモリ
リモートコピー管理情報 61
ライトシーケンス管理情報1 62_1
ライトシーケンス#カウンタ 71
721_a
ライトシーケンス管理情報エントリ (ライトシーケンス#)
721_b
721_c
721_d
721_e
721_f
721_g
ライトシーケンス管理情報エントリ領域
転送対象ライトシーケンス管理情報 73
正式化対象ライトシーケンス管理情報 74
72
ライトシーケンス管理情報(J-1) 62_(J-1)

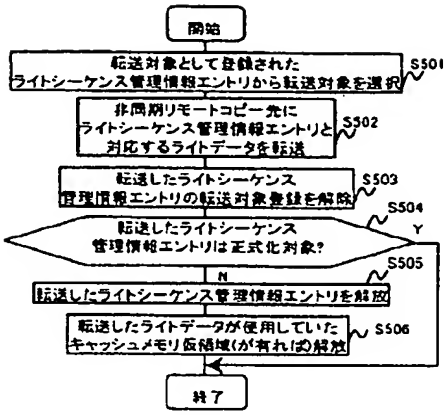
【図４】

図　4


開始
S401
ライト要求受領
S402
ライト要求元がホスト、かつ、ライト対象ボリュームはリモートコピー先でない?
Y
N
S405
ライト要求元はホスト?
Y
N
S406
ライトデータに対応させるライトシーケンス管理情報エントリを確保し、ライトシーケンス#付与
S407
ライトデータに対応させるライトシーケンス管理情報エントリを確保し、受領したライトシーケンス管理情報を格納
ライト対象ボリュームは非同期リモートコピー先?
N
Y
S408
S413
ライト対象ボリュームは非同期リモートコピー元?
N
Y
S403
ライトデータをキャッシュメモリ通常領域に格納
ライトデータをキャッシュメモリ通常領域に格納
S416
S414
非同期リモートコピー先へ未転送のライトデータをキャッシュメモリ仮領域に移設した上で、キャッシュメモリ通常領域に格納
S409
ライトデータをキャッシュメモリ仮領域に格納
ビットマップ更新処理
S417
S415
ビットマップ更新処理
ライトシーケンス管理情報エントリを正式化対象として登録
S410
ライト対象ボリュームは非同期リモートコピー元?
N
Y
S411
S412
ライトシーケンス管理情報エントリを転送対象として登録
ライト対象ボリュームは同期リモートコピー元?
N
Y
S418
S419
コピー先にライトデータとライトシーケンス管理情報を転送
ライトシーケンス管理情報エントリは転送対象でも正式化対象でもない?
N
Y
S420
S421
ライトシーケンス管理情報エントリを解放
ライト要求元にライト終了報告
S404
終了

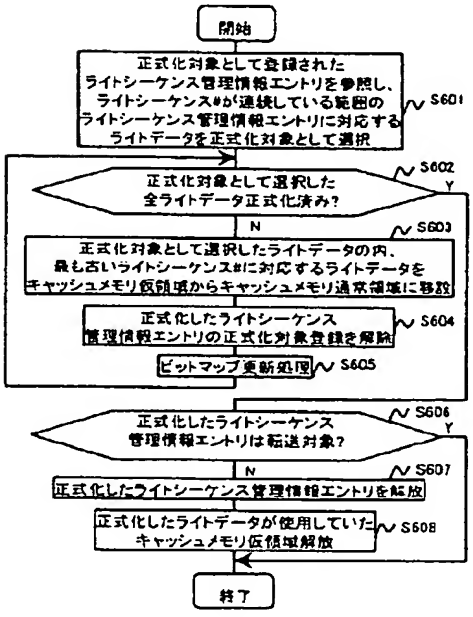
【図５】

図　５



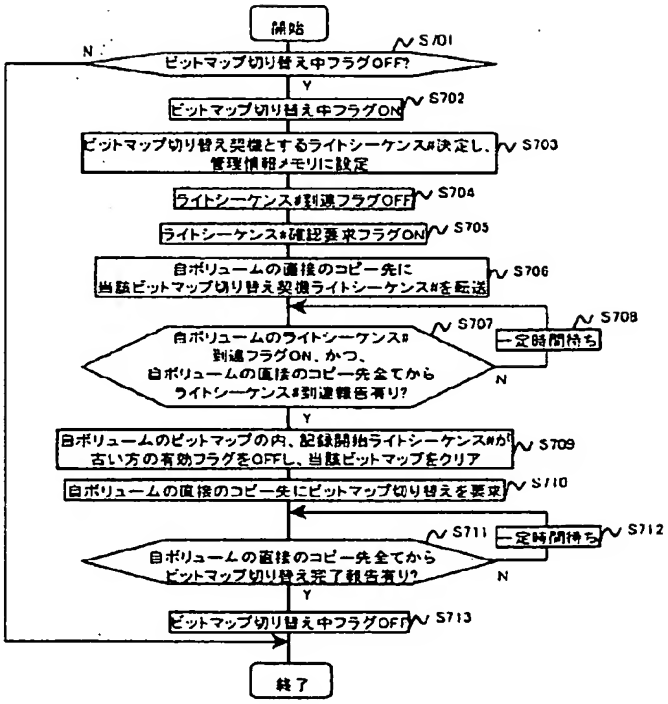
【図６】

図　６



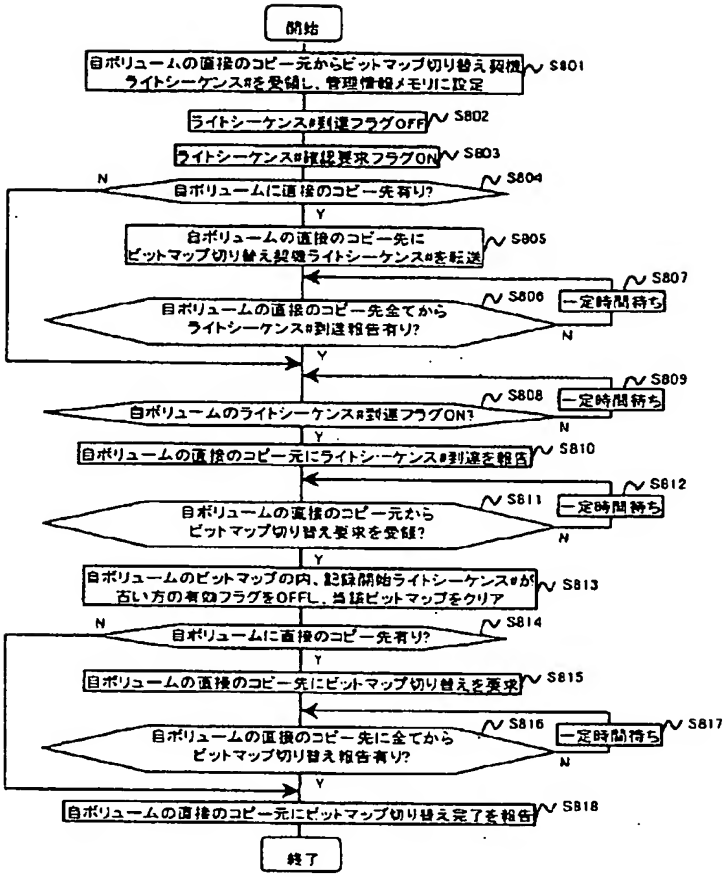
【図７】

図　７



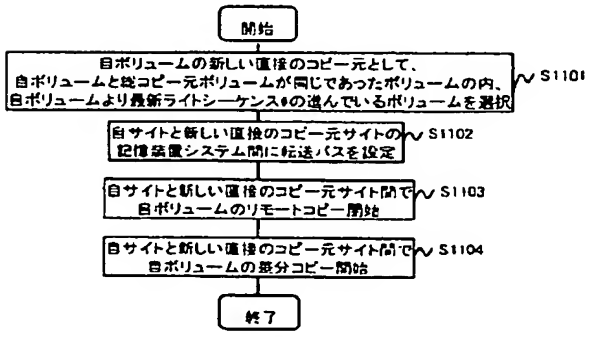
【図８】

図　8



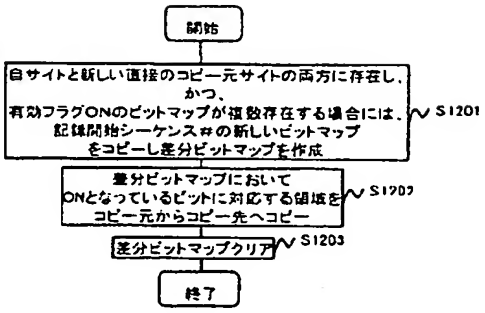
【図１１】

図　11



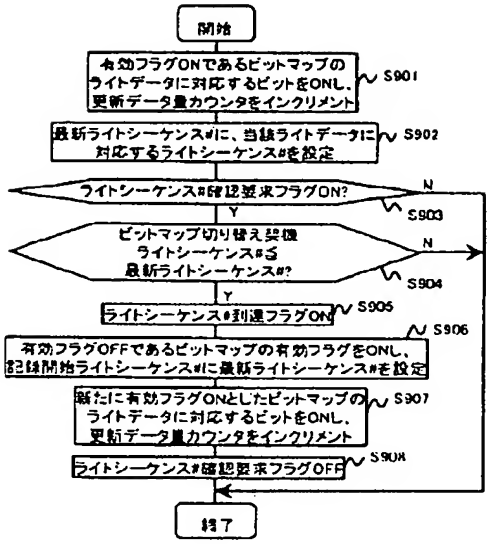
【図１２】

図　12



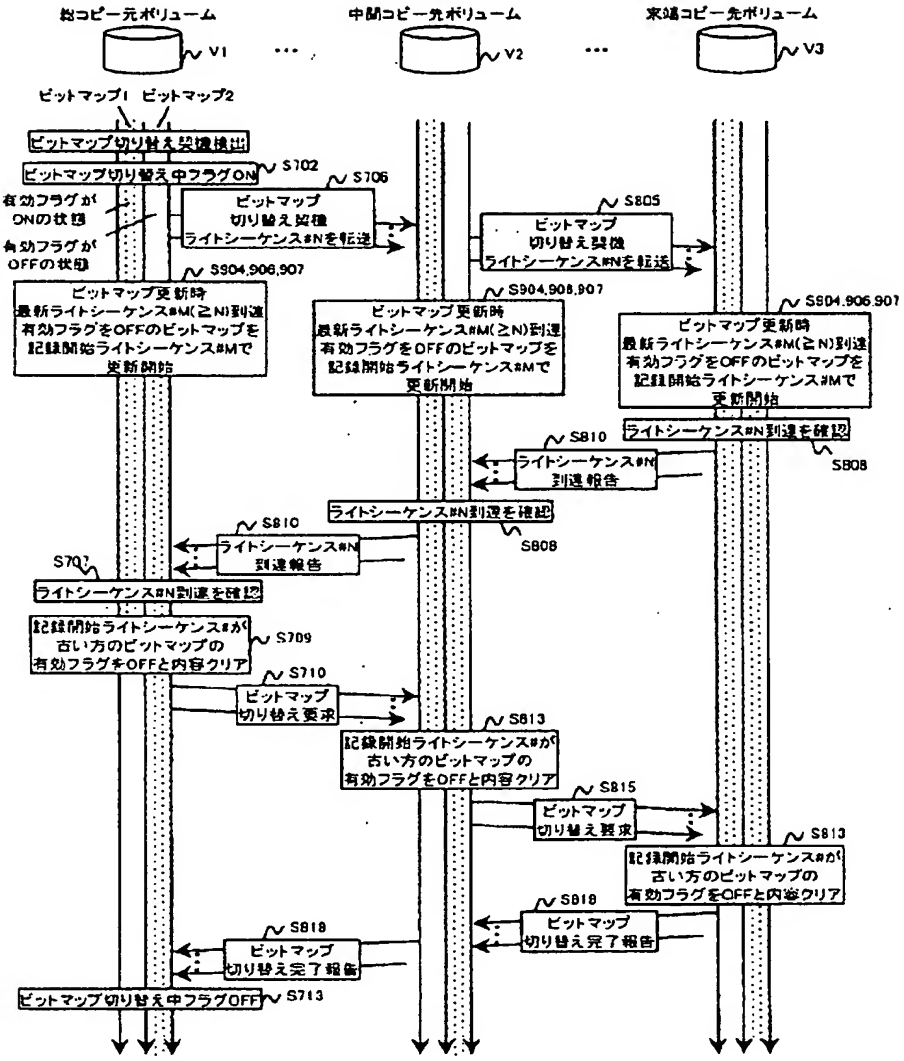
【図９】

図　９


開始
有効フラグONであるビットマップのライトデータに対応するビットをONし、更新データ量カウンタをインクリメント
S901
最新ライトシーケンス#に、当該ライトデータに対応するライトシーケンス#を設定
S902
ライトシーケンス#確認要求フラグON?
N
Y
S903
ビットマップ切り替え契機ライトシーケンス#≦最新ライトシーケンス#?
N
Y
S904
ライトシーケンス#到達フラグON
S905
有効フラグOFFであるビットマップの有効フラグをONし、記録開始ライトシーケンス#に最新ライトシーケンス#を設定
S906
新たに有効フラグONとしたビットマップのライトデータに対応するビットをONし、更新データ量カウンタをインクリメント
S907
ライトシーケンス#確認要求フラグOFF
S908
終了

【図１０】

図 10

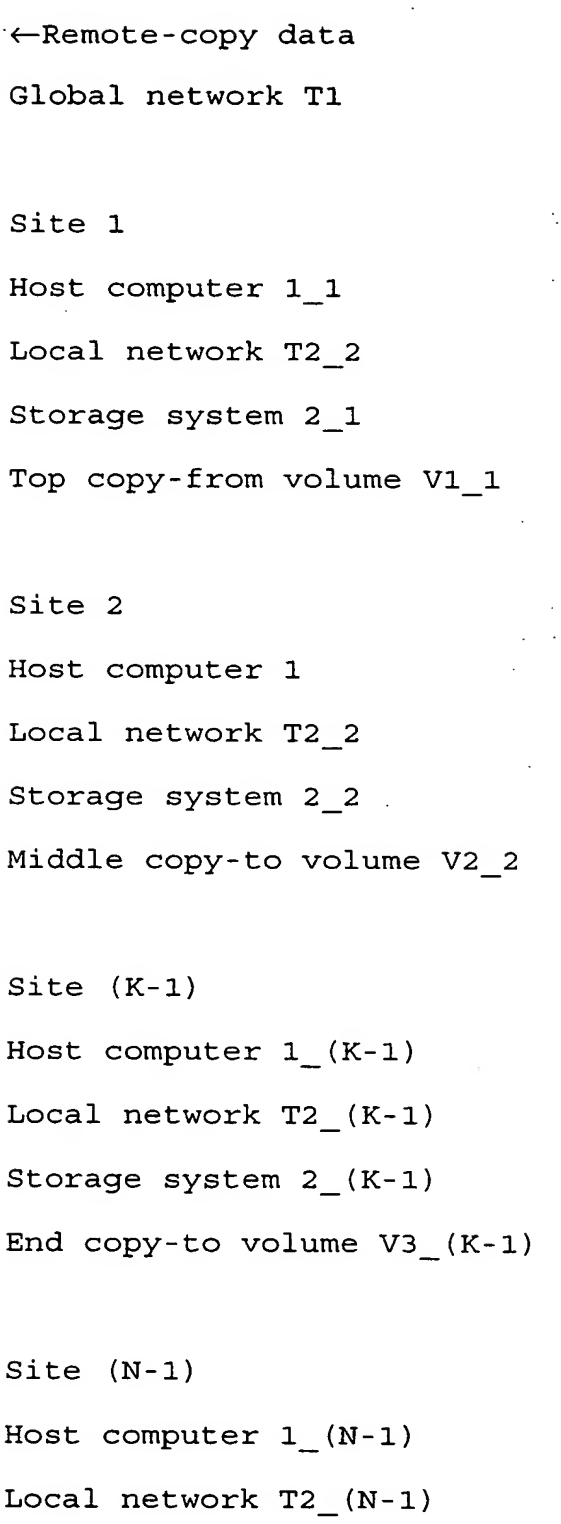


フロントページの続き

(72)発明者　檜垣　誠一
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

Ｆターム(参考)　5B065 BA01 CC08 CE06 EA24 EA35 EK05
5B082 DC09 HA03

Fig. 1


←Remote-copy data
Global network T1
Site 1
Host computer 1_1
Local network T2_2
Storage system 2_1
Top copy-from volume V1_1
Site 2
Host computer 1
Local network T2_2
Storage system 2_2
Middle copy-to volume V2_2
Site (K-1)
Host computer 1_(K-1)
Local network T2_(K-1)
Storage system 2_(K-1)
End copy-to volume V3_(K-1)
Site (N-1)
Host computer 1_(N-1)
Local network T2_(N-1)

Storage system 2_(N-1)

End copy-to volume V3_(N-1)

Site (N-2)

Host computer 1_(N-2)

Local network T2_(N-2)

Storage system 2_(N-2)

End copy-to volume V3_(N-2)

Site K

Host computer 1_K

Local network T2_K

Storage system 2_K

Middle copy-to volume V2_K

Fig. 2

2: Storage system

21: Storage controller

3: Channel interface

31: Write processing

31: Write processing

32: Bit-map update processing

32: Bit-map update processing

33: Asynchronous transmission processing

33: Asynchronous transmission processing

34: Bit-map switching start processing

34: Bit-map switching start processing

35: Bit-map switching processing

35: Bit-map switching processing

36: Difference copy processing

36: Difference copy processing

4: Disk interface

41: Asynchronous normalization processing

41: Asynchronous normalization processing

32: Bit-map update processing

32: Bit-map update processing

22: Storage device

23: Maintenance terminal

5: Cache memory

Normal area

Temporary area

6: Management information memory

61: Remote copy management information

62_1: Write-sequence management information 1

71: Write-sequence number counter

72: Write-sequence management information entry area

73: Transmission target write-sequence management information

74: Normalization target write-sequence management information

62_(J-1): Write-sequence management information (J - 1)

63_1: Volume_1 management information

801: Write-sequence management information number

802: Copy-to information

803: Copy-from information

804: Bit-map switching flag

805: Latest write sequence number

806: Bit-map switching trigger write sequence number

807: Write-sequence-number check request flag

808: Write-sequence-number reach flag

809_1: Bit map_1

810_1: Bit-map_1 validity flag

811_1: Bit-map_1 write start write sequence number

812_1: Bit-map_1 updated data amount counter

809_2: Bit map_2

810_2: Bit-map_2 validity flag

811_2: Bit-map_2 write start write sequence number

812_2: Bit-map_2 updated data amount counter

813: Difference bit map

63_(M-1) Volume number (M-1) management information

Fig. 3

2: Storage system

5: Cache memory

Normal area

Write data D0 Write data D1 Write data D2_new

Temporary area

Write data D2_old Write data D4 Write data D5

6: Management information memory

61: Remote copy management information

62_1: Write-sequence management information 1

71: Write-sequence number counter

72: Write-sequence management information entry area

721_a: Write-sequence management information entry (write-sequence number)

721_b Write-sequence management information entry (write-sequence number)

721_c Write-sequence management information entry (write-

sequence number)

721_d Write-sequence management information entry (write-sequence number)

721_e Write-sequence management information entry (write-sequence number)

721_f Write-sequence management information entry (write-sequence number)

721_g Write-sequence management information entry (no entry)

73: Transmission target write-sequence management information

74: Normalization target write-sequence management information

62_(J-1): Write-sequence management information (J-1)

Fig. 4

Start

S401: Receive a write request.

S402: Is it judged that a write requester is a host computer and a write target volume is not a remote-copy source?

S403: Store write data in a normal area of a cache memory.

S405: Is the write requester the host computer?

S406: Create a write-sequence management information entry to be associated with the write data, and give a write

sequence number to the write-sequence management information entry.

S407: Create a write-sequence management information entry to be associated with the write data, and store received write-sequence management information.

S408: Is the write target volume an asynchronous remote-copy destination?

S413: Is the write target volume an asynchronous remote-copy source?

S416: Store the write data in the normal area of the cache memory.

S417: Perform bit-map update processing.

S414: Store write data, which has not yet been transferred to an asynchronous remote-copy destination, in a temporary area of the cache memory, and then store the write data in the normal area of the cache memory.

S415: Perform bit-map update processing.

S409: Store the write data in the temporary area of the cache memory.

S410: Register the write-sequence management information entry as a target to be normalized.

S411: Is the write target volume an asynchronous remote-copy source?

S412: Register the write-sequence management information entry as a target to be transferred.

S418: Is the write target volume a synchronous remote-copy source?

S419: Transfer the write data and the write-sequence management information to a copy destination.

S420: Is it judged that the write-sequence management information entry is neither a target to be transferred nor a target to be normalized?

S421: Release the write-sequence management information entry.

S404: Notify the write requester of the write completion.

End

Fig. 5

Start

S501: Select a target to be transferred from among write-sequence management information entries registered as targets to be transferred.

S502: Transfer write data corresponding to a write-sequence management information entry to an asynchronous remote-copy destination.

S503: Delete the transferred write-sequence management information entry that has been registered as the target to be transferred.

S504: Is the transferred write-sequence management

information entry a target to be normalized?

S505: Release the transferred write-sequence management information entry.

S506: If a temporary area of a cache memory has been used by the transferred write data, release the temporary area.

End

Fig. 6

Start

S601: With reference to a write-sequence management information entry registered as a target to be normalized, select, as a target to be normalized, write data corresponding to write-sequence management information entries in a range within which write sequence numbers successively continue.

S602: Have all pieces of write data that have been selected as a target to be normalized been normalized?

S603: Among the pieces of write data that have been selected as the target to be normalized, transfer a piece of write data corresponding to the oldest write sequence number from the temporary area to the normal area in the cache memory.

S604: Delete the normalized write-sequence management information entry that has been registered as the target to be normalized.

S605: Perform bit-map update processing.

S606: Is the normalized write-sequence management information entry a target to be transferred?

S607: Release the normalized write-sequence management information entry.

S608: Release a cache memory's temporary area that has been used by the normalized write data.

End

Fig. 7

Start

S701: Is a bit-map switching flag in an OFF state?

S702: Bring the bit-map switching flag into an ON state.

S703: Determine a write sequence number used to trigger bit-map switching, and store the write sequence number in a management information memory.

S704: Bring a write-sequence-number reach flag into an OFF state.

S705: Bring a write-sequence-number check request flag into an ON state.

S706: Transfer a bit-map switching trigger write sequence number to a direct copy destination of an own volume.

S707: Is it judged that a write-sequence-number reach flag of the own volume is ON, and that write-sequence-number

reach notifications have been received from all of direct copy destinations of the own volume?

S708: Wait for a fixed period of time.

S709: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the bit map, the bit map being selected from among bit maps of the own volume.

S710: Request the direct copy destinations of the own volume to perform bit-map switching.

S711: Have bit-map switching completion notifications been received from all of the direct copy destinations of the own volume?

S712: Wait for a fixed period of time.

S713: Bring the bit-map switching flag into an OFF state.

End

Fig. 8

Start

S801: Receive a bit-map switching trigger write sequence number from a direct copy source of an own volume, and store the bit-map switching trigger write sequence number in a management information memory.

S802: Bring a write-sequence-number reach flag into an OFF state.

S803: Bring a write-sequence-number check request flag into an ON state.

S804: Does a direct copy destination exist in the own volume?

S805: Transfer a bit-map switching trigger write sequence number 806 to a direct copy destination of the own volume.

S806: Have write-sequence-number reach notifications been received from all of direct copy destinations of the own volume?

S807: Wait for a fixed period of time.

S808: Is a write-sequence-number reach flag of the own volume in an ON state?

S809: Wait for a fixed period of time.

S810: Notify the direct copy source of the own volume that the write sequence number has been reached.

S811: Has a bit-map switching request been received from the direct copy source of the own volume?

S812: Wait for a fixed period of time.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the bit map, the bit map being selected from among bit maps of the own volume.

S814: Does a direct copy destination exist in the own volume?

S815: Request the direct copy destinations of the own volume

to perform bit-map switching.

S816: Have bit-map switching completion notifications been received from all of the direct copy destinations of the own volume?

S817: Wait for a fixed period of time.

S818: Notify the direct copy source of the own volume that the bit-map switching has been completed.

End

Fig. 9

Start

S901: Bring into an ON state each bit of a bit map whose validity flag is ON, the bit corresponding to write data, and then increment a value of an updated data amount counter.

S902: Set the latest write sequence number at a write sequence number corresponding to the write data in question.

S903: Is a write-sequence-number check request flag in an ON state?

S904: Is the latest write sequence number larger than or equal to the bit-map switching trigger write sequence number?

S905: Bring the write-sequence-number reach flag into an ON state.

S906: Bring into an ON state a validity flag of a bit map

whose validity flag is OFF, and then set a write start write sequence number at a value of the latest write sequence number.

S907: Bring into an ON state each bit of a bit map whose validity flag has been newly brought into an ON state, the bit corresponding to write data, and then increment a value of the updated data amount counter.

S908: Bring the write-sequence-number check request flag in an OFF state.

End

Fig. 10

V1: Top copy-from volume

V2: Middle copy-to volume

V3: End copy-to volume

Bit map 1 Bit map 2

Detect the timing in which the bit-map switching is triggered.

S702: Bring the bit-map switching flag into an ON state.

A state in which a validity flag is ON

A state in which a validity flag is OFF

S706: Transfer a bit-map switching trigger write sequence number.

S904, S906, S907: When a bit map is updated, the latest write sequence number M is reached (≥ N), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S805: Transfer a bit-map switching trigger write sequence number N.

S904, S906, S907 When a bit map is updated, the latest write sequence number M is reached (≥ N), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S904, S906, S907: When a bit map is updated, the latest write sequence number M is reached (≥ N), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S808: Check whether or not the write sequence number N has been reached.

S810: Notify that the write sequence number N has been reached.

S808: Check whether or not the write sequence number N has been reached.

S810: Notify that the write sequence number N has been reached.

S707: Check whether or not the write sequence number N has

been reached.

S709: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S710: Request to perform bit-map switching.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S815: Request to perform bit-map switching.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S818: Notify that the bit-map switching has been completed.

S818: Notify that the bit-map switching has been completed.

S713: Bring the bit-map switching flag into an OFF state.

Fig. 11

Start

S1101: As a new direct copy source of an own volume, select a volume whose latest write sequence number is larger than that of the own volume from among volumes having the same top copy-from volume as that included in the own volume.

S1102: Provide a transmission path between a storage system in an own site and a storage system in a new direct copy-

from site.

S1103: Start remote copying between the storage system in the own site and the storage system in the new direct copy-from site.

S1104: Start difference copy processing between the storage system in the own site and the storage system in the new direct copy-from site.

End

Fig. 12

Start

S1201: Select, in a new direct copy-from site, a bit map whose validity flag is ON in both an own site and the new direct copy-from site, and if there are a plurality of bit maps each satisfying this condition, copy a bit map whose write start write sequence number is the newest to create a difference bit map.

S1202: Copy an area corresponding to bits whose value is ON in the difference bit map from a copy source to a copy destination.

S1203: Clear the difference bit map.

End

Japanese Patent Laid-open No. 2003-131917

[Title of the Invention]

Storage System

[Abstract]

[Object]

An object of the present invention is to, in remote copies made among storage controllers that are located in N sites (the number of sites is in general three or more), manage differences that are used to quickly make remote-copy data coincide with one another among the remaining sites after an arbitrary site suffers from disaster.

[Solving Means]

An update history of a remote-copy range in a site, which is directly updated from a host computer, and an update history of a remote-copy range in another site, which is a copy destination of the remote-copy range in question, are shared between the sites in question. Writing of the update histories is triggered by a certain update from the host computer. When an arbitrary site suffers from disaster, contents of the remote-copy range are made coincide by copying, according to the update history, part of contents of the remote-copy range existing between arbitrary sites that do not suffer from disaster.

[Claims]

[Claim 1]

A storage system characterized in that:

in a remote-copy configuration so formed that storage systems are connected to one another through a network among sites, said sites each including a host computer and a storage system that is connected to the host computer, and data updated by the host computer is reflected for a storage system in another site,

an update history of a remote-copy specified range of a storage system, which is directly updated by the host computer, and an update history of a copy-to remote-copy specified range of a storage system in another site, which is a remote-copy destination of the remote-copy specified range, are shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range, writing of the update histories being triggered by a certain update from the host computer, and

when an arbitrary site suffers from disaster, by copying part of contents of a remote-copy specified range or part of contents of a copy-to remote-copy specified range, which exist between arbitrary sites that do not suffer from disaster, according to the update history, the contents of the remote-copy specified range or the contents of the copy-to remote-copy specified range are made to coincide with

each other.

[Claim 2]

A storage system according to Claim 1, wherein:

areas to which the update history of the remote-copy specified range is written are provided in a multiplexed manner, and when the amount of updates from the host computer to the remote-copy specified range of the storage system, which is directly updated by the host computer, exceeds a threshold value, an update from the host computer, which triggers switching to a new update history, is determined, said determination being shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range;

in each storage system, from a point of time at which updates to the remote-copy specified range or to the copy-to remote-copy specified range reach an update from the host computer, which triggers switching to a new update history, writing to the new update history is started; and

after checking that all storage systems, each of which includes a remote-copy specified range and a copy-to remote-copy specified range, have started writing of the new update history, writing of an existing update history is interrupted, and then the existing update history is initialized.

[Claim 3]

A storage system according to Claim 2, wherein:

said determination of the update from the host computer, which triggers switching to the new update history, is made also when a length of time elapsed after the last switching of the update history exceeds a threshold value.

[Claim 4]

A storage system according to Claim 2, wherein:

exchanges of information between the storage systems are made only with a direct copy source, and a direct copy destination, corresponding to the remote-copy specified range, or the copy-to remote-copy specified range, of the storage system in question.

[Claim 5]

A storage system according to Claim 2, wherein:

exchanges of information between the storage systems are made with said information being attached to remote-copy data to be transferred or received.

[Claim 6]

A storage system according to Claim 2, wherein:

when an arbitrary site suffers from disaster, an update history to be used to make contents of a remote-copy specified range or contents of a copy-to remote-copy specified range coincide between arbitrary sites that do not suffer from disaster is selected in both of the sites on the conditions that the update history is valid, and that

updates from the host computer which has started writing are the same; and

if there are a plurality of update histories each satisfying the conditions in both of the sites, an update history whose update from the host computer which has started the writing is the latest is selected.

[Claim 7]

A storage system according to Claim 2, wherein:

if it is detected that an arbitrary site has suffered from disaster, writing of an existing update history is continued, and at a point of time when the site which has suffered from the disaster is completely recovered to a normal state, the writing of the existing update history is interrupted, and then the existing update history is initialized.

[Claim 8]

A storage system according to Claim 2, wherein:

if it is detected that an arbitrary site has suffered from disaster, by use of areas for update histories from which an existing update history is excluded, the steps described in Claim 2 are executed between sites that do not suffer from disaster.

[Detailed Description of the Invention]

[0001]

[Technical Field to which the Invention Pertains]

The present invention relates to remote-copy functions used to duplex data between storage systems with a host computer not being involved. It relates more particularly to, in remote copies made among respective storage systems located in three or more sites, when an arbitrary site suffers from disaster, if the remote copying is continued among sites that do not suffer from the disaster, how to manage contents of a remote-copy specified range so that the contents are made to coincide with each other among the storage systems.

[0002]

[Prior Art]

In order to prevent data in a storage system from being lost due to disaster or the like, the data is duplexed in another storage system located in a remote place. Examples of this duplexing function include remote copy technologies.

[0003]

In the remote copy technologies, a target volume to be remote-copied is set in a primary storage controller connected to a host computer, and also in a secondary storage controller connected to the primary storage controller; and copies are continuously being executed so that contents of a primary volume of a primary storage device connected to the primary storage controller is always

made coincide with contents of a secondary volume of a secondary storage device connected to the secondary storage controller.

[0004]

The technology for writing data in duplication between different storage controllers with a host computer not being involved is disclosed in Japanese Patent Laid-open No. 11-85408. According to the above invention, a primary storage controller, which has received from a host computer write data to which write time is added, notifies the host computer that writing of the write data has been completed, and then transfers the write time and the write data to a secondary storage controller in the order of the write time. In a secondary storage controller, the write time and the write data which have been received from the primary storage controller are stored in a nonvolatile cache memory to ensure that data stored before this specified write time is reliable.

[0005]

The remote-copy operation is roughly classified into two: a synchronous remote copy for transferring write data to the secondary storage controller before notifying the host computer of the completion of a write request; an asynchronous remote copy for transferring write data to the secondary storage controller in asynchronization with the

write request after notifying the host computer of the completion of the write request. If a remote-copy function is used for applications such as a database, it is necessary to update a secondary volume in the same order as the update order in which a primary volume has been updated so as to ensure the update order of a log. In order to achieve the above, in the case of the asynchronous remote copy, every time a write is made from a host computer, a sequence number is added to write data in a primary storage controller. Then, the write is reflected in a secondary volume according to the sequence number in a secondary storage controller. [0006]

In the conventional remote-copy operation described above, the remote-copy operation may be temporarily interrupted (for example, in a case where the remote-copy operation is interrupt by a link failure between the primary storage controller and the secondary storage controller, or the like). In such temporal interruption, with the objective of making contents of the secondary volume coincide with those of the primary volume as quickly as possible when the remote-copy operation is restarted, an update history may be written in the primary storage controller as difference information between the primary and secondary volumes. In the update history, updates from the host computer to the primary volume performed during the

interruption of the remote-copy operation are recorded. After that, when the remote-copy operation is restarted, by using the written update history to reflect the contents of the primary volume in the secondary volume, it is possible to make the contents of the secondary volume coincide with those of the primary volume.

[0007]

The difference information in the conventional remote copy usually means update locations from the host computer. More specifically, a location at which the primary volume is updated but the secondary volume is not updated is written. Therefore, the difference information is in general stored on the primary storage controller side.

[0008]

In recent years, with the increase in speed of networks, and with the decrease in cost of networks, costs of data transmission between remote locations are decreasing. For this reason, to increase the amount of transferred data in the remote-copy operation, and thereby to further increase the availability of remote-copy data, there is a demand that remote copies are made among N sites (the number of sites is three or more) so that data is kept duplexed even if one site suffers from disaster.

[0009]

As operational management of remote copies among N

sites, if one site among N sites suffers from disaster, it is desirable that a remote-copy configuration be made again by use of the remaining (N - 1) sites so as to continue the remote-copy operation. In this case, in order to continue the remote-copy operation, it is necessary to quickly make remote-copy data coincide between the remaining (N - 1) sites. In other words, it is desirable to perform difference management among (N - 1) sites so that the remote-copy data is quickly made coincide with one another by copying the differences.

[0010]

It is because if the difference management is not performed, there may arise a case where it becomes necessary to copy all remote-copy data from a copy source to a copy destination. In comparison with the copy of the differences, the copy of all remote-copy data takes a long time, and the availability of data is decreased while the copy is being executed. Therefore, it is not possible to make full use of the redundancy achieved by the number of sites.

[0011]

Here, as an example, on the assumptions that a remote-copy configuration is made with three sites, and that the conventional difference management method is applied to remote-copy operation, problems which arise are considered. Here, three sites are designated as A, B, C.

[0012]

First of all, if a flow of remote-copy data to be transferred among the sites are A → B → C, in the event that the site B suffers from disaster, it is desirable that differences be copied between a volume in the site A and a volume in the site C before restarting the remote-copy operation. Because the difference information in the conventional remote copy is in general stored on the primary storage controller side, difference management information between the site A and the site B is stored in the site A, and difference management information between the site B and the site C is stored in the site B. For this reason, if the site B suffers from disaster, the difference management information between the site B and the site C is lost. Accordingly, differences between the site A and the site C cannot be known.

[0013]

In addition, if a flow of remote-copy data to be transferred among the sites is A → B and A → C, in the event that the site A suffers from disaster, it is desirable that differences be copied between a volume in the site B and a volume in the site C before restarting the remote-copy operation. In this case, because difference management information between the site A and the site B and that between the site A and the site C are stored in the site A

that has suffered from the disaster, differences between the site B and the site C cannot be known.

[0014]

In order to cope with this problem, a method is conceivable in which difference management is performed beforehand between sites between which remote-copy data is not exchanged during the normal operation. In the above example, a difference between the site A and the site C is managed before the site B suffers from disaster; and a difference between the site B and the site C is managed before the site A suffers from disaster. However, if this method is used, the amount of difference information between sites which must be managed beforehand increases in proportion to the number of sites. Furthermore, it is not possible to manage differences in the event of unexpected circumstances including, for example, a case where two sites concurrently suffer from disaster.

[0015]

[Problems to be Solved by the Invention]

The conventional difference management method used for remote copies between two sites does not take difference management among N sites into consideration.

[0016]

Therefore, when remote copies are being made among N sites, if one site suffers from disaster, and consequently,

the remote-copy operation is continued among the remaining (N - 1) sites that do not suffer from disaster, there arises a problem in that a remote-copy configuration can be made again only after transferring all remote-copy data from a new copy-from site to a new copy-to site.

[0017]

In addition, if the conventional difference management method used for remote copies between two sites is applied beforehand to a pair of new sites when a remote-copy configuration is made again after a site has suffered from disaster, there arises a problem in that the amount of difference management information of each site increases with the increase in the number of sites. Further, if a plurality of sites suffer from disaster, there arises another problem in that differences cannot be managed.

[0018]

An object of the present invention is to, in remote copies made among storage systems that are located in N sites (the number of sites is in general three or more), after an arbitrary site suffers from disaster, manage differences that are used to quickly make remote-copy data coincide with each other among the remaining sites. Another object of the present invention is to keep the amount of information used to manage the differences constant without depending on the number of sites N among which remote copies

are made.

[0019]

[Means for Solving the Problems]

In order to achieve the above-mentioned objects, according to one aspect of the present invention, there is provided a storage system, wherein:

in a remote-copy configuration so formed that storage systems are connected to one another through a network among sites, said sites each including a host computer and a storage system that is connected to the host computer, and data updated by the host computer is reflected for a storage system in another site,

an update history of a remote-copy specified range of a storage system, which is directly updated by the host computer, and an update history of a copy-to remote-copy specified range of a storage system in another site, which is a remote-copy destination of the remote-copy specified range, are shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range, writing of the update histories being triggered by a certain update from the host computer, and

when an arbitrary site suffers from disaster, by copying part of contents of a remote-copy specified range or part of contents of a copy-to remote-copy specified range, which exist between arbitrary sites that do not suffer from

disaster, according to the update history, the contents of the remote-copy specified range or the contents of the copy-to remote-copy specified range are made to coincide with each other.

[0020]

In addition, areas to which the update history of the remote-copy specified range is written are provided in a multiplexed manner, and when the amount of updates from the host computer to the remote-copy specified range of the storage system, which is directly updated by the host computer, exceeds a threshold value, an update from the host computer, which triggers switching to a new update history, is determined, the determination being shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range; in each storage system, from a point of time at which updates to the remote-copy specified range or to the copy-to remote-copy specified range reach an update from the host computer, which triggers switching to a new update history, writing to the new update history is started; and after checking that all storage systems, each of which includes a remote-copy specified range and a copy-to remote-copy specified range, have started writing of the new update history, writing of an existing update history is interrupted, and then the existing update history is initialized.

[0021]

In addition, the determination of the update from the host computer, which triggers switching to the new update history, may be made also when a length of time elapsed after the last switching of the update history exceeds a threshold value.

[0022]

In addition, exchanges of information between the storage systems may also be made only with a direct copy source, and a direct copy destination, corresponding to the remote-copy specified range, or the copy-to remote-copy specified range, of the storage system in question.

[0023]

Further, the exchanges of information between the storage systems may also be made with said information being attached to remote-copy data to be transferred or received.

[0024]

Moreover, when an arbitrary site suffers from disaster, an update history to be used to make contents of a remote-copy specified range or contents of a copy-to remote-copy specified range coincide between sites that do not suffer from disaster is selected in both of the sites on the conditions that the update history is valid, and that updates from the host computer which has started writing are the same; and additionally, if there are a plurality of

update histories each satisfying the conditions in both of the sites, an update history whose update from the host computer which has started the writing is the latest may also be selected.

[0025]

Furthermore, an instruction which is used in the event of a failure in a remote-copy destination is given from a maintenance terminal, the instruction instructing to allow or disallow the execution of a remote copy to the other remote-copy destinations for each remote-copy destination of a remote-copy specified range or of a copy-to remote-copy specified range; and when a failure occurs in a remote-copy destination, remote copies from the storage system in question to the remote-copy destinations may be controlled according to the instruction.

[0026]

[Modes for Carrying out the Invention]

A first embodiment of the present invention will be described with reference to drawings below.

[0027]

Fig. 1 is a diagram illustrating how remote copies among N sites are made. In each site, each of a plurality of host computers 1, each of a plurality of storage systems 2, and the like, are connected to one another through each local network T2 such as a LAN (Local Area Network) or a SAN

(Storage Area Network). On the other hand, among the storage systems 2 in those sites, storage systems 2 between which a remote copy is made are connected to each other through a global network T1. In general, the global network T1 is a public telephone service, which is often leased from a communication service provider with charge. However, the present invention is not limited by configurations of the local networks T2, and a configuration of the global network T1.

[0028]

Target volumes to be remote-copied are roughly classified into: a top copy-from volume V1 that does not have a copy-from volume, and that has a copy-to volume; a middle copy-to volume V2 that has a copy-from volume and a copy-to volume; and an end copy-to volume V3 that has a copy-from volume, and that does not have a copy-to volume. The relationship among the target volumes to be remote-copied forms a tree shape with the top copy-from volume V1 being a vertex. Data to be remote-copied is transferred in succession from the top copy-from volume V1 to the end copy-to volume V3 through the middle copy-to volume V2.

[0029]

Fig. 1 illustrates a configuration in which a volume in the site 1 is specified as the top copy-from volume V1. In this configuration, a remote copy is made through the

middle copy-to volume V2 in the site 2 to the end copy-to volume V3 in the site (K - 1) and the end copy-to volume V3 in the site N, and also a remote copy is made through the middle copy-to volume V2 in the site K to the end copy-to volume V3 in the site (N -1).

[0030]

Fig. 2 is a diagram illustrating a configuration of the storage system 2. A storage controller 21 comprises components of: channel interfaces 3 each of which is connected to the host computer 1; disk interfaces 4 each of which is connected to a storage device 22; duplexed nonvolatile management information memories 6 for storing management information; and duplexed nonvolatile cache memories 5 for storing data. The above components are connected to one another through paths. Each of the channel interfaces 3 and each of the disk interfaces 4 are connected to the duplexed management information memories 6 and also to the duplexed cache memories 5. In addition, the storage controller 21 further comprises a maintenance terminal 23 used to instruct the storage controller 21 in question, and to display an internal state of the storage controller 21 in question.

[0031]

The channel interface 3 comprises: write processing 31 for handling a write request received from the host

computer 1, and a write request for a remote copy received from another storage system 2 that is a source of the remote copy; bit-map update processing 32 for updating a bit map corresponding to a range within which a write has been made; asynchronous transmission processing 33 for, if the own storage system 2 is a source of an asynchronous remote copy, transferring write data to said another storage system 2 as a copy destination in asynchronization with a write request; bit-map switching start processing 34 for triggering, from the storage system 2 having the top copy-from volume V1, switching of a bit map 809 covering all volumes that are copy destinations of the volume in question; bit-map switching processing 35 for receiving a bit-map switching request from the bit-map switching start processing 34, and then for switching the bit map; and difference copy processing 36 for, when reconfiguring a remote-copy configuration among remaining sites after suffering from disaster in an arbitrary site, making volume contents in each of the remaining sites coincident with those in the other remaining sites.

[0032]

The disk interface 4 comprises: asynchronous normalization processing 41 for, if the own storage system 2 is a destination of an asynchronous remote copy, normalizing remote-copy data that is temporarily saved as temporary data

at the time of writing from a copy source; and bit-map update processing 32 that is the same as the bit-map update processing 32 on the side of the channel interface 3.

[0033]

There are a plurality of pieces of processing for each kind of the above processing. As a unit of existence, the bit-map switching start processing 34, the bit-map switching processing 35, and the difference copy processing 36 may exist on a volume basis; or the asynchronous transmission processing 33 and the asynchronous normalization processing 41 may also exist for each write-sequence management information 62 described below.

[0034]

The cache memories 5 store data that is stored in the storage device 22, and that is read/written from/to the host computer 1. Each of the cache memories 5 can be broadly classified into a normal area and a temporary area. The reason why each cache memory 5 has the temporary area is that if the own storage system 2 is a source of an asynchronous remote copy, when untransferred remote-copy data is written to a copy destination again, it is necessary to save untransferred data before the write, or that if the own storage system 2 is a destination of an asynchronous remote copy, it is necessary to save remote-copy data existing after the remote-copy data is received from a copy

source before the remote-copy data is normalized by the asynchronous normalization processing 41.

[0035]

The management information memories 6 store management information required for the operation of the storage system 2. The management information includes remote-copy management information 61. The remote-copy management information 61 includes write-sequence management information 62 and volume management information 63.

[0036]

The write-sequence management information 62 is used to normalize remote-copy data in the write order received from the host computer 1 in an asynchronous remote copy. There are a plurality of pieces of the write-sequence management information 62. Each piece of write-sequence management information 62 includes: a write-sequence number counter 71 for writing the write order received from the host computer 1; a write-sequence management information entry 721 of a write-sequence management information entry area 72, which is assigned on a write data basis, and in which a write sequence number is saved so that the write sequence number is associated with write data; a transmission target write-sequence management information 73 for, in a copy source of an asynchronous remote copy, registering untransferred remote-copy data in a copy

destination; and normalization target write-sequence management information 74 for, in a copy destination of an asynchronous remote copy, registering remote-copy data that has been received from a copy source but has not yet been normalized.

[0037]

Fig. 3 is a diagram illustrating a structure of the write-sequence management information 62 and the relationship between the write-sequence management information 62 and the cache memory 5. The write-sequence management information entry 721 is assigned to each piece of write data. A location of the write data in the cache memory 5 is stored in the write-sequence management information entry 721. In addition, the transmission target write-sequence management information 73 and the normalization target write-sequence management information 74 may also be configured to have a queue structure for connecting the write-sequence management information entry 721.

[0038]

In addition, if it is necessary to ensure the write order received from the host computer 1 among a plurality of volumes, a write sequence number is assigned to write data in each of the plurality of volumes by use of the write-sequence number counter 71 of the same write-sequence

management information 62, and then the write sequence numbers are registered in the same transmission target write-sequence management information 73. In a copy destination, the write sequence numbers are registered in the same normalization target write-sequence management information 74, and normalizations are made according to the write sequence numbers.

[0039]

Returning to Fig. 2, the volume management information 63 is management information about a target volume to be remote-copied. There are a plurality of pieces of the volume management information 63. Each piece of volume management information 63 comprises: a write-sequence management information number 801 for specifying write-sequence management information 62 used by an own volume; copy-to information 802 of the volume in question; copy-from information 803 of the volume in question; a bit-map switching flag 804 used to exclude bit-map switching: the latest write sequence number 805 used when the volume in question is a copy-to volume; a bit-map switching trigger write sequence number 806 which is compared with the latest write sequence number so as to trigger bit-map switching; a write-sequence-number check request flag 807 which is used to request at the time of bit-map switching to check whether or not the latest write sequence number 805 reaches the bit-

map switching trigger write sequence number 806 in the own storage system 2; a write-sequence-number reach flag 808 for indicating that the latest write sequence number 805 has reached the bit-map switching trigger write sequence number 806; a plurality of bit maps 809, each of which stores updated points of the volume in question; a plurality of bit-map validity flags 810, each of which indicates that a corresponding bit map is valid; a plurality of bit-map write start write sequence numbers 811 each indicating the write sequence number from which writing to a corresponding bit map is started; a plurality of bit-map updated data amount counters 812 each indicating the amount of updated data of a corresponding bit map; and a difference bit map 813 which is used when it becomes necessary to make contents of a volume coincide with those in the other sites as a result of suffering from disaster.

[0040]

The copy-to information 802 and the copy-from information 803 include: information that is required to communicate with the copy-from or copy-to storage system 2; and information for identifying a copy-from or copy-to volume.

[0041]

According to the embodiment of the present invention, for the sake of simplification, management information is

provided on a volume basis (the volume management information 63). However, this does not limit the scope of the present invention. The management information can be provided in an arbitrary unit that can be shared between the storage systems 2. Accordingly, it is possible to make a remote copy in the arbitrary unit that can be shared between the storage systems 2. The write processing 31 which is executed in each storage system 2 will be described with reference to Fig. 4.

[0042]

Here, a write requester and classification of write data, which are targeted by the write processing 31, will be outlined so that the description can be easily understood. The write requester is the host computer 1, or the storage system 2 that is a source of a remote copy.

[0043]

Depending on a write target volume, write data from the host computer 1 can be classified into: 1) write data having no remote-copy destination; and write data having a remote-copy destination. The write data having a remote-copy destination can be further classified into: 2) write data having only a synchronous remote-copy destination; 3) write data having only an asynchronous remote-copy destination; and 4) write data having both a synchronous remote-copy destination and an asynchronous remote-copy

destination.

[0044]

Depending on a write target volume, write data from the storage system 2 which is a remote-copy source can be classified into: write data that is written by a synchronous remote copy; and write data that is written by an asynchronous remote copy. As is the case with the above 1) through 4), the write data that is written by a synchronous remote copy can be classified into: 5) write data having no remote-copy destination; and write data having a remote-copy destination. The write data having a remote-copy destination can be further classified into: 6) write data having only a synchronous remote-copy destination; 7) write data having only an asynchronous remote-copy destination; and 8) write data having both a synchronous remote-copy destination and an asynchronous remote-copy destination. In addition, as is the case with the above 1), 4), the write data which is written by an asynchronous remote copy can be classified into: 9) write data having no remote-copy destination; and 10) write data having only an asynchronous remote-copy destination. The reason why 2) and 3) are excluded from the targets is that in the synchronous remote copy, a write completion notification is transferred to the host computer on the completion of a remote copy. Accordingly, even if, in the middle copy-to volume V2,

remote-copy data received by an asynchronous remote copy is remote-copied to a volume in another site by a synchronous remote copy, it is a meaningless and unrealistic matter.

[0045]

In the write processing 31, first of all, a write request is received (step S401), and then a judgment is made as to whether or not conditions that a write requester is the host computer 1, and that a write target volume is not a remote-copy source, hold true (step S402).

[0046]

If both of the conditions in the step S402 hold true, the above write request is a write request from the host computer 1 to a volume that does not relate to a remote copy (more specifically, write data is equivalent to the 1) described above). Accordingly, write data is stored in a normal area of the cache memory 5 (step S403), and then the write requester is notified of write completion (step S404). The storage controller 21 writes this write data to the storage device 22 in asynchronization with the write request from the host computer 1.

[0047]

If one or both of the conditions in the step S402 does not hold true, a check is made as to whether or not the write requester is the host computer 1 (step S405).

[0048]

If the condition in the step S405 holds true (more specifically, the write data is equivalent to any of 2), 3), 4) described above), the write-sequence management information entry 721 to be associated with the write data is created, and then a write sequence number is given to the write-sequence management information entry 721. To be more specific, with reference to the write-sequence management information number 801 of the volume management information 63 corresponding to a volume to which a write has been made from the host computer 1, the write-sequence management information entry 721 of the write-sequence management information 62 corresponding to the write-sequence management information number 801 is created. After that, a value of the write-sequence number counter 71 is set, and then the value of the write-sequence number counter 71 is added so that the write-sequence management information entry 721 is associated with the write data.

[0049]

The addition of the write sequence number is executed regardless of synchronization/asynchronization of the remote copy. This is because even if direct copy destinations of the volume in question include only synchronous remote-copy destinations, a further destination of these copy destinations may execute an asynchronous remote copy.

[0050]

If the condition in the step S405 does not hold true (more specifically, the write data is equivalent to any of 6) through 10) described above), the write-sequence management information entry 721 to be associated with the write data is created, and then the write-sequence management information 62 including a write sequence number, which has been received together with the write data, is stored (step S407).

[0051]

Next, the write data is stored in the cache memory 5. A method for storing the write data in the cache memory 5 depends on a write target volume. There are three cases as follows: A) a case where the write target volume is an asynchronous remote-copy destination (more specifically, equivalent to any of 9), 10) described above); B) a case where the write target volume is not an asynchronous remote-copy destination but an asynchronous remote-copy source (more specifically, equivalent to any of 4), 7), 8) described above); and C) a case other than the above A), B) (more specifically, equivalent to any of 2), 5), 6) described above). This is because in the case of A) it is necessary to temporarily store received write data, and then formally write the write data according to the write sequence number. This is also because in the case of B) if untransferred write data is written to an asynchronous

remote-copy destination, it is necessary to independently store this write data until old write data is transferred.

[0052]

First of all, a judgment is made as to whether or not a write target volume is an asynchronous remote-copy destination (step S408).

[0053]

If the condition in the step S408 holds true (more specifically, equivalent to A) described above), it is necessary to write the write data in question according to a write sequence number indicating the update order from the host computer 1. Accordingly, the write data is temporarily stored in the temporary area of the cache memory 5 (step S409). Then, the write-sequence management information entry 721 corresponding to the write data is registered in the normalization target write-sequence management information 74 as target data to be normalized (step S410). In Fig. 3, write-sequence management information entries 721_e, 721_f correspond to the target data to be normalized.

[0054]

Moreover, a judgment is made as to whether or not a write target volume is a copy source of an asynchronous remote copy (step S411). If the condition in the step S411 holds true (more specifically, equivalent to any of 3), 4), 7), 8), 10) described above), the write-sequence management

information entry 721 corresponding to the write data is registered in the transmission target write-sequence management information 73 as target data to be transferred (step S412). In Fig. 3, write-sequence management information entries 721_b, 721_c, 721_d, 721_e correspond to the target data to be transferred.

[0055]

If the condition in the step S408 holds true, a further judgment is made as to whether or not the write target volume is a copy source of an asynchronous remote copy (step S413).

[0056]

If the condition in the step S413 holds true (more specifically, equivalent to B) described above), write data that has not yet been transferred to an asynchronous remote-copy destination is stored in the temporary area of the cache memory 5, and then the write data is stored in the normal area of the cache memory 5 (step S414). In Fig. 3, the write-sequence management information entries 721_b, 721_c, 721_d correspond to the write data. Further, the bit-map update processing 32 is executed (step S415). The bit-map update processing 32 will be described later. After that, the process proceeds to the step 411.

[0057]

If the condition in the step S413 does not hold true

(more specifically, equivalent to C) described above), the write data in question is neither a copy destination of an asynchronous remote copy nor a copy source of the asynchronous remote copy. Accordingly, the write data is stored in the normal area of the cache memory 5 (step S416). In Fig. 3, write-sequence management information entries 721_a correspond to the write data. Further, the bit-map update processing 32 is executed (step S417). The bit-map update processing 32 will be described later. After that, the process proceeds to the step 418.

[0058]

Next, a judgment is made as to whether or not the write target volume is a copy source of a synchronous remote copy (step S418). If the condition in the step S418 holds true (more specifically, equivalent to any of 2), 4), 6), 8) described above), write data and contents of the write-sequence management information entry 721 are transferred to a copy destination of the synchronous remote copy (step S419). With the object of shortening the time taken after receiving a write request before a write completion notification is transferred to the write requester, the synchronous remote copy in question may also be executed in parallel with respect to a plurality of copy destinations.

[0059]

Next, a judgment is made as to whether the write-

sequence management information 721 corresponding to the write data in question is registered in neither the transmission target write-sequence management information 73 nor the normalization target write-sequence management information 74 (step S420). If the condition in the step S420 holds true (more specifically, equivalent to any of 2), 5), 6) described above), the write-sequence management information entry 721 corresponding to the write data in question is not necessary. Therefore, the write-sequence management information entry 721 is released (step S421), which corresponds to a write-sequence management information entry 721_g in Fig. 3.

[0060]

Lastly, a write completion notification is transferred to the write requester (step S404), before the processing ends.

[0061]

The asynchronous transmission processing 33 which is executed in a storage system 2 having a copy-from volume of an asynchronous remote copy among the storage systems 2 will be described with reference to Fig. 5. This processing may also be executed for each write-sequence management information 62.

[0062]

First of all, a target to be transferred is selected

from among the write-sequence management information entries 721 registered in the transmission target write-sequence management information 73 (step S501).

[0063]

Next, the selected write-sequence management information entry 721 and write data corresponding to the write-sequence management information entry 721 in question are transferred to an asynchronous remote-copy destination (step S502).

[0064]

Next, the transferred write-sequence management information entry 721 is deleted from the transmission target write-sequence management information 73 (step S503).

[0065]

Next, a judgment is made as to whether or not the transferred write-sequence management information entry 721 is registered also in the normalization target write-sequence management information 74 (step S504). If the condition in the step S504 does not hold true, management as target data of the asynchronous remote copy becomes unnecessary. Accordingly, the transferred write-sequence management information entry 721 is released (step S505). Here, if the transferred write data has been stored in the temporary area of the cache memory 5, the temporary area in question is also released (step S506).

[0066]

The asynchronous normalization processing 41 which is executed in a storage system 2 having a copy-to volume of an asynchronous remote copy among the storage systems 2 will be described with reference to Fig. 6. This processing may also be executed for each write-sequence management information 62.

[0067]

First of all, with reference to the write-sequence management information entry 721 registered in the normalization target write-sequence management information 74, a range within which write sequence numbers successively continue is selected as a target to be normalized (step S601).

[0068]

Next, a judgment is made as to whether or not all write data which is the selected target to be normalized has already been normalized (step S602).

[0069]

If the condition in the step S602 holds true, the process proceeds to a step S606.

[0070]

If the condition in the step S602 does not hold true, among the pieces of write data which have been selected as the target to be normalized, a piece of write data

corresponding to the oldest write sequence number is transferred from the temporary area to the normal area in the cache memory 5, and then the piece of write data is normalized there (step S603).

[0071]

Next, the normalized write-sequence management information entry 721 is deleted from the normalization target write-sequence management information 74 (step S604). Next, the bit-map update processing 32 is executed (step S605). The bit-map update processing 32 will be described later. After that, the process returns to the step S602.

[0072]

Next, a judgment is made as to whether or not the normalized write-sequence management information entry 721 is registered also in the transmission target write-sequence management information 73 (step S606). If the condition in the step S606 does not hold true, management as target data of the asynchronous remote copy becomes unnecessary. Accordingly, the normalized write-sequence management information entry 721 is released (step S607). In addition, the temporary area of the cache memory 5, in which the normalized write data has been stored, is released (step S608).

[0073]

The bit-map switching start processing 34 which is

executed in a storage system 2 having the top copy-from volume V1 among the storage systems 2 will be described with reference to Fig. 7. This processing may also be executed on a volume basis if a value of an updated data amount counter 812 of the bit map 809 whose bit-map validity flag 810 is in an ON state exceeds a fixed threshold value (if the number of bit maps 809 whose bit-map validity flag 810 is in an ON state is two or more, a bit map whose write start write sequence number 811 is the oldest is selected from among them).

[0074]

In addition, if the elapsed time after the lastly executed bit map switching exceeds a fixed threshold value, the bit-map switching start processing 34 may also be executed on a volume basis.

[0075]

First of all, for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1, in order to prevent bit map switching from being newly started before the last bit map switching ends, a judgment is made as to whether or not the bit-map switching flag 804 is in an OFF state (step S701).

[0076]

If the condition in the step S701 does not hold true,

the bit map switching is still being executed. Therefore, the processing is ended.

[0077]

If the condition in the step S701 holds true, the bit-map switching flag 804 is brought into an ON state (step S702).

[0078]

Next, a write sequence number used to trigger bit-map switching is determined. Then, the bit-map switching trigger write sequence number 806 of the management information memory is set at the determined write sequence number (step S703). A method for determining a write sequence number used to trigger bit-map switching may be the following: A write sequence number is so generated that the latest write sequence number 805 relating to the middle copy-to volume V2 and the end copy-to volume V3 does not exceed the generated write sequence number before a notification is transferred from the top copy-from volume V1 to all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and further to all end copy-to volumes V3 corresponding to the top copy-from volume V1. This write sequence number is used as the bit-map switching trigger write sequence number 806.

[0079]

In addition, it is intended that the bit-map

switching trigger write sequence number 806 becomes a write start write sequence number of a bit map 809 to which bit-map switching is made. However, if write sequences of a plurality of volumes are managed with a piece of write-sequence management information 62, the determined bit-map switching trigger write sequence number 806 is not always assigned to write data of the volume in question. However, a write sequence number that is first assigned to the volume in question with the determined bit-map switching trigger write sequence number 806 being exceeded is uniquely determined for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1. Therefore, the write sequence number that is first assigned to the volume in question with the bit-map switching trigger write sequence number 806 being exceeded can be used as a write start write sequence number of the bit map 809 to which bit-map switching is made.
[0080]

Next, the write-sequence-number reach flag 808 is brought into an OFF state (step S704). This write-sequence-number reach flag 808 is used to indicate that the latest write sequence number 805 becomes larger than or equal to the bit-map switching trigger write sequence number 806.
[0081]

Next, the write-sequence-number check request flag 807 is brought into an ON state (step S705). This write-sequence-number check request flag 807 is used to request the bit-map update processing 32 to compare the latest write sequence number 805 with the bit-map switching trigger write sequence number 806.

[0082]

Next, the bit-map switching trigger write sequence number 806 is transferred to all direct copy destinations of the volume in question to request a write-sequence-number reach notification when the latest write sequence number exceeds the bit-map switching trigger write sequence number 806 (step S706).

[0083]

Next, judgments are made as to whether or not the write-sequence-number reach flag 808 of the volume in question is ON, and whether or not write-sequence-number reach notifications are received from all of the direct copy destinations of the volume in question (step S707). If one or both of the conditions in the step S707 does not hold true, after waiting for a fixed period of time (step S708), the judgments in the step S707 are made again. It is to be noted that notifications from the direct copy destinations of the volume in question are successively stored in the copy-to information 802.

[0084]

If both of the conditions in the step S707 hold true, a validity flag 810 of a bit map 809 corresponding to the oldest write start write sequence number 811 is brought into an OFF state, said bit map 809 being selected from among the bit maps 809 in the own site. Then, the bit map 809 in question is cleared (step S709). This is because at this point of time, for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1, the latest write sequence numbers 805 of these volumes exceed the bit-map switching trigger write sequence number 806, and consequently writing to a new bit map 809 has already been started.

[0085]

Next, all of the direct copy destinations of the volume in question are requested to perform bit map switching (step S710).

[0086]

Next, a judgment is made as to whether or not bit map switching completion notifications are received from all of the direct copy destinations of the volume in question (step S711). If the condition in the step S711 does not hold true, after waiting for a fixed period of time (step S712), the judgment in the step S711 is made again.

[0087]

If the condition in the step S711 holds true, the bit map switching is completed. Therefore, the bit-map switching flag 804 is brought into an OFF state (step S713), before the processing ends.

[0088]

The bit-map switching processing 35 which is executed in a storage system 2 having the middle copy-to volume V2 and the end copy-to volume V3 among the storage systems 2 will be described with reference to Fig. 8.

[0089]

First of all, the bit-map switching trigger write sequence number 806 is received from a copy-from site. Then, the bit-map switching trigger write sequence number 806 of the middle copy-to volume V2, and that of the end copy-to volume V3, are set at a value of the received bit-map switching trigger write sequence number 806 (step S801). Here, the middle copy-to volume V2 and the end copy-to volume V3 are targets of bit map switching.

[0090]

Next, the write-sequence-number reach flag 808 is brought into an OFF state (step S802), and the write-sequence-number check request flag 807 is brought into an ON state (step S803).

[0091]

Next, a judgment is made as to whether or not a direct copy destination exists in the volume in question (step S804).

[0092]

If the condition in the step S804 does not hold true (more specifically, the volume in question is equivalent to the end copy-to volume V3), the process proceeds to a step S808.

[0093]

If the condition in the step S804 holds true (more specifically, the volume in question is equivalent to the middle copy-to volume V2), the bit-map switching trigger write sequence number 806 is transferred to all direct copy destinations of the volume in question to request a write-sequence-number reach notification when the latest write sequence number 805 exceeds the bit-map switching trigger write sequence number 806 (step S805).

[0094]

Next, a judgment is made as to whether or not write-sequence-number reach notifications are received from all of the direct copy destinations of the volume in question (step S806). If the condition in the step S806 does not hold true, after waiting for a fixed period of time (step S807), the judgment in the step S806 is made again.

[0095]

If the condition in the step S806 holds true, a judgment is made as to whether or not the write-sequence-number reach flag of the volume in question is ON (step S808). If the condition in the step S808 does not hold true, after waiting for a fixed period of time (step S809), the judgment in the step S808 is made again.

[0096]

If the condition in the step S808 holds true, the direct copy source of the volume in question is notified that the write sequence number has been reached (step S810).

[0097]

Next, a judgment is made as to whether or not a bit-map switching request has been received from a direct copy source of the volume in question (step S811). If the condition in the step S811 does not hold true, after waiting for a fixed period of time (step S812), the judgment in the step S811 is made again.

[0098]

If the condition in the step S811 holds true, the bit-map validity flag 810 of a bit map 809 corresponding to the oldest write start write sequence number 811 is brought into an OFF state, the bit map 809 being selected from among the bit maps 809 of the volume in question. Then, the bit map 809 in question is cleared (step S813).

[0099]

Next, a judgment is made as to whether or not a direct copy destination exists in the volume in question (step S814).

[0100]

If the condition in the step S814 does not hold true (more specifically, the volume in question is equivalent to the end copy-to volume V3), the process proceeds to a step S818.

[0101]

If the condition in the step S814 holds true (more specifically, the volume in question is equivalent to the middle copy-to volume V2), all of the direct copy destinations of the volume in question are requested to perform bit-map switching (step S815).

[0102]

Next, a judgment is made as to whether or not bit map switching completion notifications are received from all of the direct copy destinations of the volume in question (step S816). If the condition in the step S816 does not hold true, after waiting for a fixed period of time (step S817), the judgment in the step S816 is made again.

[0103]

If the condition in the step S816 holds true, the direct copy source of the volume in question is notified that the bit map switching has been completed (step S818),

and then the processing ends.

[0104]

The bit-map update processing 32, which is executed in the write processing 31 and the asynchronous normalization processing 41, will be described with reference to Fig. 9.

[0105]

First of all, each bit of a bit map 809 whose bit-map validity flag 810 is ON is brought into an ON state, the said bit corresponding to write data, and then the updated data amount counter 812 corresponding to the updated bit map 809 is incremented (step S901).

[0106]

Next, the latest write sequence number 805 is updated to a write sequence number corresponding to the write data in question (step S902).

[0107]

Next, a judgment is made as to whether or not the write-sequence-number check request flag 807 is ON (step S903).

[0108]

If the condition in the step S903 does not hold true, the processing ends.

[0109]

If the condition in the step S903 holds true, a

judgment is made as to whether or not the latest write sequence number 805 is larger than or equal to the bit-map switching trigger write sequence number 806 (step S904).

[0110]

If the condition in the step S904 does not hold true, the processing ends.

[0111]

If the condition in the step S904 holds true, the write-sequence-number reach flag 808 is brought into an ON state (step S905).

[0112]

Next, the bit-map validity flag 810 of a bit map 809 whose bit-map validity flag 810 is OFF is brought into an ON state, and then the write start write sequence number 811 is set at a value of the latest write sequence number 805 (step S906).

[0113]

Next, each bit of the bit map 809 whose bit-map validity flag 810 has been newly brought into the ON state is switched to ON, the said bit corresponding to write data, and then the updated data amount counter 812 corresponding to the updated bit map 809 is incremented (step S907).

[0114]

Lastly, the write-sequence-number check request flag 807 is brought into an OFF state (step S908), before ending

the processing.

[0115]

Fig. 10 is a diagram schematically illustrating processing in each site, and communications between sites, at the time of bit-map switching.

[0116]

The bit-map switching start processing 34 is executed in a site having the top copy-from volume V1; and the bit-map switching processing 35 is executed in a site having the middle copy-to volume V2 or the end copy-to volume V3. However, the bit-map update processing 32 is partially included. An appropriate step number is given to each step to be executed.

[0117]

Fig. 10 schematically illustrates bit-map switching from a bit map 1 to a bit map 2 in the whole site. After that, the bit-map switching is alternately executed from the bit map 2 to the bit map 1, and further from the bit map 1 to the bit map 2.

[0118]

By shifting the bit-map switching as the whole site from a state in which a single bit map is updated, through a state in which double bit maps are temporarily updated, to a state in which a single bit map is updated again, the write start write sequence number 811 can always share the same

bit map 809 between arbitrary sites during an arbitrary moment including the bit-map switching, and thereby volume contents can be made to coincide with those in the site in which the write sequence number goes ahead.

[0119]

In addition, in the event that a failure is detected in a certain site during the bit-map switching, transfer of the bit-map switching trigger write sequence number 806 to the storage system 2 in each site may fail. In this case, in the storage system 2 having the top copy-from volume V1, the bit map 809 is not cleared in the step S709. Then, in a step S710, by requesting storage systems 2 in the other sites to transfer a bit-map switching completion notification without clearing the bit map 809, it is possible to continue management of differences including the site where the failure has occurred.

[0120]

Moreover, if each storage system 2 has triple bit maps 809 and a failure is detected in a certain site, as described above, by preventing the bit map 809 used when the failure has been detected from being cleared, and by continuing the bit-map switching by use of remaining bit maps, it is possible to continue the management of the differences including the site where the failure has occurred, and thereby to prevent the differences between

normal sites from increasing. Even if the number of sites in which a failure has occurred increases, by keeping unchanged the bit map 809, which is prevented from being cleared, until all sites become normal, it is possible to continue the management of the differences including the site where the failure has occurred.

[0121]

In addition, by transferring and receiving usual remote-copy data together with the contents of communications between two storage systems performed at the time of the bit-map switching, it is possible to reduce a communication load between the sites.

[0122]

Steps of making the volume contents coincide between sites after one of the sites suffers from disaster will be described with reference to Fig. 11.

[0123]

First of all, as a new direct copy source of an own volume that has lost a direct copy source as a result of the disaster from which the site has suffered, a volume whose latest write sequence number 805 is larger than that of the own volume is selected from among volumes having the same top copy-from volume V1 as that included in the own volume (step S1101).

[0124]

Next, a transfer path is provided between the storage system 2 in an own site and the storage system 2 in a new direct copy-from site (step S1102).

[0125]

Next, remote copying is started between the storage system 2 in the own site and the storage system 2 in the new direct copy-from site (step S1103). Thereafter, updates made from the new direct copy-from site to the copy-from volume are successively reflected in a copy-to volume in the own site.

[0126]

Lastly, the difference copy processing 36 between the storage system 2 in the own site and the storage system 2 in the new direct copy-from site is started (step S1104). The difference copy processing 36 will be described later.

[0127]

By continuing remote copying also after completion of the difference copy processing 36, the remote copying becomes usual remote copy operation.

[0128]

As the new direct copy-from site described above, a site on the upstream side whose distance to the top copy-from volume V1 is closer than the distance from the own site to the top copy-from volume V1 may also be registered beforehand for the own volume.

[0129]

The difference copy processing 36 which is one of the steps of making the volume contents coincide between sites after one of the sites suffers from disaster will be described with reference to Fig. 12.

[0130]

First of all, a bit map 809 whose bit-map validity flag 810 is ON in both an own site and a new direct copy-from site is selected in the new direct copy-from site. Here, if there are a plurality of bit maps each satisfying this condition, a bit map 809 whose write start write sequence number 811 is the newest is selected from among the bit maps. Then, contents of the selected bit map 809 is copied to the difference bit map 813 that is used to determine data to be copied in the difference copy processing 36 (step S1201). The reason why the bit map 809 is copied to the difference bit map 813 is the following: In the new direct copy-from site, the remote copying may be continued even after the site has suffered from disaster, and as a result, the bit map 809 whose bit-map validity flag 810 is ON may be continuously updated. This makes it necessary to save the differences between current volumes in both sites so as to prevent the amount of differences to be copied from increasing.

[0131]

Next, part corresponding to bits whose value is ON in the difference bit map 813 is copied from a copy-from volume in the new direct copy-from site to a copy-to volume in the own site (step S1202).

[0132]

Lastly, the difference bit map 813 is cleared (step S1203), and then the processing ends.

[0133]

[Effects of the Invention]

According to the present invention, in remote copies made among storage controllers that are located in N sites (the number of sites is in general three or more), after one site suffers from disaster, it is possible to manage differences that are used to quickly make remote-copy data coincide with each other among the remaining sites. In addition, it is possible to keep the amount of information used to manage the differences constant without depending on the number of sites N among which remote copies are made.

[Brief Description of the Drawings]

[Fig. 1]

Fig. 1 is a diagram illustrating how remote copies among N sites are made according to an embodiment of the present invention by way of example.

[Fig. 2]

Fig. 2 is a diagram illustrating a configuration of a

storage system according to an embodiment of the present invention by way of example.

[Fig. 3]

Fig. 3 is a diagram illustrating the relationship between write-sequence management information and a cache memory according to an embodiment of the present invention by way of example.

[Fig. 4]

Fig. 4 is a flowchart illustrating an example of write processing that is executed in a storage system according to an embodiment of the present invention.

[Fig. 5]

Fig. 5 is a flowchart illustrating an example of asynchronous transmission processing that is executed in a storage system of an asynchronous remote copy source according to an embodiment of the present invention.

[Fig. 6]

Fig. 6 is a flowchart illustrating an example of asynchronous normalization processing that is executed in a storage system of an asynchronous remote copy destination according to an embodiment of the present invention.

[Fig. 7]

Fig. 7 is a flowchart illustrating an example of bit-map switching start processing that is executed in a storage system having a top copy-from volume V1 according to an

embodiment of the present invention.

[Fig. 8]

Fig. 8 is a flowchart illustrating an example of bit-map switching processing that is executed in a storage system having a middle copy-from volume/an end copy-from volume according to an embodiment of the present invention.

[Fig. 9]

Fig. 9 is a flowchart illustrating an example of bit-map update processing which is executed in write processing and asynchronous normalization processing according to an embodiment of the present invention.

[Fig. 10]

Fig. 10 is a diagram schematically illustrating bit-map switching according to an embodiment of the present invention.

[Fig. 11]

Fig. 11 is a flowchart illustrating steps of making the volume contents coincide between sites after one of the sites suffers from disaster according to an embodiment of the present invention by way of example.

[Fig. 12]

Fig. 12 is a flowchart illustrating an example of difference copy processing which is one of the steps of making the volume contents coincide between sites after one of the sites suffers from disaster according to an

embodiment of the present invention.

[Explanations of Symbols and Reference Numerals]

1 ... Host computer, 2 ... Storage system, 21 ... Storage controller, 22 ... Storage device, 23 ... Maintenance terminal, 3 ... Channel interface, 31 ... Write processing, 32 ... Bit-map update processing, 33 ... Asynchronous transmission processing, 34 ... Bit-map switching start processing, 35 ... Bit-map switching processing, 36 ... Difference copy processing, 4 ... Disk interface, 41 ... Asynchronous normalization processing, 5 ... Cache memory, 6 ... Management information memory, 62 ... Write-sequence management information, 63 ... Volume management information, 71 ... Write-sequence number counter, 721 ... Write-sequence management information entry, 73 ... Transmission target write-sequence management information, 74 ... Normalization target write-sequence management information, 804 ... Bit-map switching flag, 805 ... Latest write sequence number, 806 ... Bit-map switching trigger write sequence number, 807 ... Write-sequence-number check request flag, 808 ... Write-sequence-number reach flag, 809 ... Bit map, 810 ... Bit-map validity flag, 811 ... Bit-map write start write sequence number, 812 ... Bit-map updated data amount counter, 813 ... Difference bit map, T1 ... Global network, T2 ... Local network, V1 ... Top copy-from volume, V2 ... Middle copy-to volume, V3 ... End

copy-to volume.